Routledge
Taylor & Francis Group

Language Assessment Quarterly

ISSN: 1543-4303 (Print) 1543-4311 (Online) Journal homepage: https://www.tandfonline.com/loi/hlaq20

Specifying Progression in Academic Speaking: A Keyword Analysis of CEFR-Based Proficiency Descriptors

Armin Berger

To cite this article: Armin Berger (2020) Specifying Progression in Academic Speaking: A Keyword Analysis of CEFR-Based Proficiency Descriptors, Language Assessment Quarterly, 17:1, 85-99, DOI: 10.1080/15434303.2019.1689981

To link to this article: https://doi.org/10.1080/15434303.2019.1689981

Published online: 28 Nov 2019.

LANGUAGE ASSESSMENT QUARTERLY
2020, VOL. 17, NO. 1, 85–99
https://doi.org/10.1080/15434303.2019.1689981

Routledge
Taylor & Francis Group

Specifying Progression in Academic Speaking: A Keyword Analysis of CEFR-Based Proficiency Descriptors

Armin Berger

University of Vienna, Vienna, Austria

ABSTRACT

The Common European Framework of Reference (CEFR), with its illustrative scales and salient features of spoken language at the reference levels, is widely used as the base for rating scales for performance testing. If practitioners want to measure and report even small gains in proficiency, they need to adapt the descriptors to their local context and make further subdivisions. However, progression from one band to the next is often defined intuitively without much empirical support. This article presents an attempt to create a finer, empirically-based differentiation for academic speaking at C1 and C2 in the form of common reference points, which specify progression in a way that is more precise than the level descriptions in the CEFR but not too specific to lose their referential nature. To validate the progression, the common reference points are compared to the results of a keyword analysis of C1 and C2 descriptors for speaking from several CEFR-related sources. The results appear to confirm the soundness of the suggested progression in general terms. The findings reflect both the potential and the limitations of a keyword analysis for present purposes, indicating that the approach taken is complementary to other forms of validation.

Operationalizing constructs in rating scales is an ongoing challenge in language test development, especially the definition of progression and its theoretical and empirical underpinning continues to be a key issue of concern. In practice, the Common European Framework of Reference for Languages or CEFR (Council of Europe, 2001) is often used as the base for operational definitions of progression, along with the more recent CEFR Companion Volume with New Descriptors (Council of Europe, 2018). However, although the CEFR provides a measurement-based description of increasing language proficiency, it is not designed to function as or to contain ready-made rating instruments (Fulcher, 2016). Instead, the CEFR descriptors have to be extended and adapted to suit the needs of the specific assessment context (Saville, 2012). Supplementary descriptors and sublevels are a desideratum, particularly in tertiary language education, where the specifications for C1 and C2 are often too vague and generic.

As much as the illustrative scales and the salient characteristics that emerge across the different CEFR levels (Council of Europe, 2001, 2018) are intended to assist in the formulation of assessment criteria and levels of attainment (Council of Europe, 2001), in practice, there is little empirical guidance on how to differentiate within the levels. This is particularly the case at the upper end of the proficiency range, which, unlike levels A2 to B2, does not have so-called 'plus levels'. While a number of projects have provided additional details on the horizontal dimension of the CEFR's common reference levels by specifying or enriching the categories for describing communicative activities and aspects of competence (Green, 2012; Hawkins & Filipović, 2012; North, Ortega, & Sheehan, 2010), far less attention has been paid to the vertical dimension, i.e. a more fine-grained

CONTACT Armin Berger armin.berger@univie.ac.at Department of English and American Studies, University of Vienna, Spitalgasse 2-4, Hof 8.3, 1090 Wien, Austria

description of the progress in proficiency in those categories. In order to refine progression, local scale developers therefore often resort to abstract descriptor formulations using qualifiers like 'almost always', 'generally', 'somewhat', or 'generally not'. Such semantically formulated scales are notoriously vague and open to interpretation, however. What would aid scale developers greatly is an empirical specification of progression which distinguishes sublevels in real and concrete terms.

This article aims to present such a specification for academic speaking at levels C1 and C2, which grew out of a pedagogical need in an Austrian university context to define progression in speaking proficiency within the C levels in a nuanced way. The progression is based on a set of analytic rating scales for the assessment of oral presentations and interactions developed by the Austrian University English Language Teaching and Testing (ELTT) initiative, a working group consisting of linguists and language teachers at the Universities of Graz, Klagenfurt, Salzburg, and Vienna. While the rating scale development cum validation process and its specific methodology has been presented elsewhere (Berger, 2015, 2018), the focus here is on the salient characteristics of the progression, i.e. common reference points at five band levels abstracted from the specific ELTT descriptors across the scale categories. Such finer distinctions at the upper proficiency levels satisfy the need in many higher education contexts to capture and report even small gains in proficiency.

The specific purpose of the study was to validate the progression as shown in the locally-developed common reference points. To this end, a keyword analysis was conducted to extract distinctive features from a number of other CEFR-based descriptors for speaking at levels C1 and C2. These keywords were then compared to the common reference points, and similarities between the keywords and the common reference points were interpreted as confirming the soundness of the progression. While keyword analysis has been widely used in discourse and genre analysis, for example to obtain descriptive accounts of a particular text type, it has not yet been fully exploited to investigate the properties of proficiency descriptors.

To provide the wider context, the article starts with some general remarks about progressions for the purposes of language assessment. Then it goes on to outline the genesis of the common reference points. The subsequent sections present the results of the keyword analysis and discuss them in relation to the locally-developed reference points. Finally, a brief evaluation of the keywords approach for present purposes rounds off the paper.

Defining progression in language proficiency

In language education, the term 'progression' embodies a multitude of overlapping concepts. Many course curricula and the stages of attainment associated with them are in effect progressions with holistic level descriptors (Adey, 1997). In a different but related sense, the term can refer to learning progressions, which map out specific (empirically validated) learning pathways or a particular sequence of knowledge and skills that students are expected to acquire as they progress through a course or program (Bailey & Heritage, 2014). Within language assessment, in particular, progressions are commonly associated with proficiency scales, usually in the form of examination levels, proficiency frameworks or rating scales. While proficiency frameworks, such as the CEFR, serve as a context-neutral and unifying point of reference which integrates different purposes, users and qualifications, and to which local progressions can be aligned (Jones & Saville, 2008), rating scales are operational definitions of a construct with context-specific proficiency descriptors against which a leaner's performance is compared (Fulcher, 2012). Depending on the purpose of the rating scale, the progression can have different functions. Alderson's (1991) well-known classification of scales into three types provides a useful heuristic for identifying the functions of such progressions. A *user-oriented* progression conveys information about typical behavior at the various levels, for example to help learners understand the incremental growth from less to more proficient; an *assessor-oriented* progression assists raters in distinguishing a performance at one level reliably from performances at adjacent levels; and a *constructor-oriented* progression guides test developers in selecting tasks as well

as language features at the right level. In the present study, the focus is on progression in relation to scales and frameworks oriented towards assessors and scale constructors.

Defining progression and assigning particular features to particular levels is one of the central challenges in scale construction (Luoma, 2004). Typically, progression is operationalized in two ways: either systematically or in terms of criterial features. The "systematic approach" (North, 2014, p. 26) attempts to define progression in a consistent, seamless fashion, mentioning every feature recurrently over the full range of the scale. The distinctions are usually indicated by adverbs of degree or frequency, such as 'most', 'many' and 'some'. While this way of capturing variation has the face validity of being objective, North (2014) reminds us that, in fact, the opposite is the case as differences are reduced to purely semantic variation which is meaningful only in relation to some internalized understanding of the relevant standard. Alternatively, and desirably, descriptors capture what is significant and criterial only at the level concerned. In the so-called "salient features approach" (North, 2014, p. 26), descriptors represent real and concrete as opposed to semantic differences, and the descriptions are cumulative in that learners at a higher level are, as a matter of course, expected to demonstrate the abilities described at the lower levels of the scale as well.

What is considered salient and criterial in rating scales is often a matter of judgement. Less frequently, such decisions are the result of empirical approaches grounded in measurement theory (e.g. North, 2000) or performance data (e.g. Fulcher, Davidson, & Kemp, 2011). From a second language acquisition perspective, critics call for such features to be firmly based on research into developmental patterns studied over a longer period of time; otherwise, the proficiency levels may bear little relation to how language ability actually develops (Hulstijn, 2007). Although more research on the relationship between developmental stages and levels of second-language proficiency is underway (Hulstijn, Alderson, & Schoonen, 2010), North (2014) points out that such sequences of acquisition are still not available in a form that would allow scale developers to use them as a basis for scale construction. Accordingly, Jones and Saville (2008) argue that scale developers cannot and need not wait for such a theoretical basis, desirable though it may be. Scales should be regarded as "heuristic models of reality" (Jones & Saville, 2008, pp. 497–498), as opposed to statements about reality itself, and the primary aim should be to integrate useful aspects of theory eclectically into a practical validity model for test construction.

In practice, the CEFR's common reference levels often serve as the point of departure for defining progressions, to the extent possible given the political and content-related criticisms surrounding the CEFR (Deygers, Zeidler, Vilcu, & Carlsen, 2017). The levels have a horizontal and a vertical dimension. The former refers to a range of communicative activities, strategies and language competences, and the latter to an ascending sequence of levels representing progress in those categories. Language learners develop along both dimensions, with more proficient users being able to perform an ever-increasing number of communicative activities in increasingly complex and sophisticated ways. Together with the descriptive scheme, the reference levels are intended, among other things, to provide a source for the definition of assessment criteria and the formulation of levels of attainment (Council of Europe, 2001).

What the CEFR does not provide and, by nature, does not intend to do so, is ready-made rating scales for operational use. It has been pointed out multiple times that, not least because of the language-neutral nature of the CEFR, the reference levels are underspecified, lacking in detail with regard to key features that are characteristic and indicative of a particular proficiency level (Hawkins & Filipović, 2012; Milanovic, 2009; North, 2014). This underspecification and incompleteness is partly deliberate and partly accidental. Milanovic (2009) reminds us that the CEFR was created expressly as a *common framework* for language learning, teaching and assessment intended to be multi-purpose, flexible, open and dynamic in nature and orientation. What follows is that practitioners need to adjust the framework to suit their local purposes. On the other hand, the acknowledged underspecification is in part the result of how the CEFR came into being. After qualitative and quantitative validation, some gaps appeared along the proficiency continuum as descriptors had to be rejected owing to quality control issues. This problem was particularly pertinent at the C levels

(North, 2014) so that further elaboration is required. One of the most recent and systematic endeavors to extend the CEFR is presented in the Council of Europe's (2018) CEFR Companion Volume with New Descriptors. However, much as the changes and additions represent a considerable improvement of the C levels, the supplementary descriptors do not obviate the need in many contexts for further subdivisions of the broad levels to measure and report even minimal progress.

While a considerable amount of research has been conducted in relation to what is criterial at different CEFR levels, less attention has been paid to the progression within the levels. North et al. (2010), for example, created an inventory of functions, grammar, discourse markers, vocabulary areas and topics for levels A1 to C1. The various projects in the English Profile Programme, in particular, aim to characterize the language that learners of English typically use at each level of the CEFR, specifically in relation to vocabulary and grammar (Capel, 2015; Hawkins & Filipović, 2012; O'Keeffe & Mark, 2017). Of relevance to this study, Green (2012) used a keywords approach to identify criterial language functions at C1 and C2 by analyzing pedagogical and assessment materials judged to be linked to the CEFR levels. While all this research, despite its limitations (O'Sullivan, 2014), has made a significant contribution to defining the linguistic content at different levels, few studies have addressed the vertical dimension in terms of a more fine-grained description of the progress within each level, particularly at the C levels.

The study reported here is one such endeavor to attend to the vertical dimension at the C levels. It validates a progression of academic speaking based on a number of rating scale descriptors for academic presentations and interactions. While many validation studies of this kind attempt to replicate the intended order or hierarchy of the original scale descriptors by means of different scaling methodologies (e.g. Berger, 2015; Kaftandjieva & Takala, 2002), this one exploits a keyword analysis to identify key concepts in other C-level descriptors for speaking and then checks whether the proposed progression is in line with these key concepts. Keyword analysis is often utilized to determine the aboutness or style of a text, to obtain descriptive accounts of specific text types or to detect types of discourse or ideology (Archer, Culpeper, & Rayson, 2009; Baker, 2009; Scott & Tribble, 2006). In this study, a keyword analysis is used to highlight salient features of proficiency levels. The main advantage for present purposes is that it helps researchers to recognize important concepts in proficiency descriptors with less bias than would result from an inductive inference (Baker, 2004). Thus, methodologically, the study is related to Green's (2012) approach to exploring criterial differences in functionality by subjecting a range of C-level materials to a keyword analysis. It is different, however, in that it applies the keyword analysis exclusively to proficiency descriptors. Before the focus of this article shifts to the keyword analysis, the following section provides some more information on the context of the study; it describes the genesis of the common reference points in the course of validating the ELTT scales.

Context of the study

Aiming to professionalize assessment practices in Austrian university English departments, a team of language teachers and linguists created a set of analytic rating scales for the certification of speaking proficiency in English as a foreign language at the end of the language competence courses in their BA programs. Taken together, the scales comprise 174 descriptors in six categories: *lexico-grammatical resources and fluency, pronunciation and vocal impact, structure and content (presentations), genre-specific presentation skills, content and relevance (interactions)* and *interaction skills*. Although the team had developed the scales in a systematic and principled manner (see Berger & Heaney, 2018 for details), a number of concerns remained, most notably that the progression from one band to the next as prescribed by the descriptors lacked empirical validation.

To investigate the continuum of increasing speaking ability underlying the scales, a multi-method validation study was conducted in several stages (Berger, 2015, 2018). First, in a sorting task, 21 experienced language teachers from five Austrian universities were asked to reconstruct the intended

order of the descriptors. Second, the sorting task data was subjected to a multi-faceted Rasch analysis (Linacre, 2013). Third, eight expert raters linked the descriptors to 153 video performances, producing a total of 21,909 data points which were analyzed by means of multi-faceted Rasch measurement as well. The most stable descriptors across all procedures were then reintegrated into coherent proficiency descriptions in five bands. Finally, the performance descriptions in each band of the revised scales were examined for their common ground. The assumption was that if commonalities exist between different categories in the same band and if they can be abstracted from the specific descriptors through an inductive inference, such features could serve as common reference points characterizing that band. To this end, the calibrated descriptors were grouped according to trait categories and listed in descending order of their logit values, similar to the methodology used by North (2000). The juxtaposition of the different categories facilitated the inspection of the content coherence of the different bands, i.e. the identification of salient features the descriptors seemed to have in common. The progression across the various scale categories showed a fairly coherent picture, suggesting that the bands have good content integrity. The CEFR (Council of Europe, 2001, 2018) provides such a description of the content coherence for spoken language from A1 to C2, including the narrower "plus levels" for "basic" and "independent" language use, but it does not provide a more fine-grained set of distinctions regarding "proficient" language use. The suggested reference points, presented in descending order of proficiency in Figure 1, may help to fill this gap. The progression as reflected in the common reference points is the focus of the current study. For a fuller understanding of the relationship between the descriptors and the common reference points, the interested reader is referred to Berger (2015, pp. 184–203, 236–254, 285–290).

Typical and indicative of the band level concerned, the common reference points are intended to guide future scale development in a coherent and transparent way should the team wish to extend or modify the scales at a later stage. However, the locally-developed common reference points could also be useful beyond the Austrian university context, because although the ELTT scales were devised for a particular purpose, the more abstract reference points offer possible criterion statements for other settings. They are more specific than the CEFR's current level characterizations at C1 and C2, but not too specific to be divested of their referential function, which allows scale developers in other contexts to adapt these statements for their own purposes and yet refer back to a framework linked to the CEFR. Signposting progression in a nuanced yet generic way, the common reference points could provide benchmarks for similar scale development projects in tertiary contexts where there is a need for further subdivisions without losing the reference to the CEFR's C levels. They could also inform course curricula, teaching and learning objectives, materials design as well as other pedagogical purposes in tertiary language education which rely on the ability to observe even minor progress at an advanced level.

There are, however, two possible threats to the wider applicability of the common reference points: Firstly, they are based on a relatively small sample of descriptors related to academic presentations and outcome-oriented discussions. Secondly, all teachers who took part in the validation study were based at Austrian universities, so the progression may, in the worst case, reflect little more than a specific teaching culture in relation to a given student population. What had not taken place yet is a validation of the locally-developed common reference points themselves to see if they can potentially have currency in other contexts as well. The present study sought to do just that.

The keyword analysis

Methodology

In order to confirm or disconfirm, at least in general terms, the pattern of increasing speaking proficiency, C1 and C2 descriptors from a number of CEFR-based documents were surveyed. Relating the common reference points back to CEFR-linked descriptors from other contexts would reveal similarities and discrepancies, supporting or challenging the validity of the ELTT

progression. Accordingly, the purpose of this study was to compare the progression as reflected in the common reference points with established knowledge as expressed through other proficiency descriptors associated with spoken production and interaction at C1 and C2. The main research question was whether the basic pattern of increasing speaking proficiency as shown in the locally-developed common reference points is reflected in other descriptors for speaking at C1 and C2 from different contexts.

The process began with the collection of proficiency descriptors for speaking from various sources, including the CEFR Companion Volume with New Descriptors (Council of Europe, 2018), European Language Portfolio (ELP) models and other CEFR-related proficiency scales. For conceptual and practical reasons, a number of potentially instructive sources were excluded at the outset. Modern textbooks used in tertiary language programs, for example, were of little avail because they often have a very specific EAP focus without any explicit link to the CEFR. Although intriguing, descriptors without any connection to the CEFR would have obscured the comparison between the keywords and the common reference points. Some textbooks are associated with a particular CEFR level, albeit in a global, intuitive rather than any more definitive manner; none of the textbooks surveyed for this analysis provided any detailed information about the nature of their relationship to the CEFR, let alone meta-level information in the form of can-do statements. Furthermore, sources like test specifications, examination handbooks and rating scales are largely absent from the collection. Much as testing instruments used by tertiary institutions for assessing their students' speaking proficiency could have been informative, especially because they could have provided a deeper level of granularity allowing a direct comparison with the ELTT bands, such materials are normally confidential and thus difficult to obtain. In addition, scale descriptors covering a relatively narrow proficiency range are usually interdependent, with distinctions between scale bands relying for the most part on qualifiers like 'some', 'many' or 'most' rather than real and meaningful distinctions. Such relative descriptors would have been of little use for the purpose of identifying concrete distinctions. Table 1 provides an overview of the sources from which descriptors were extracted.

A descriptor was selected if it satisfied the following criteria: (a) it was expressly related to spoken production or interaction, or to any other category in the ELTT scales; (b) it was considered relevant to speaking for academic purposes; (c) there was an explicit link to levels C1 or C2; (d) the descriptor represented a stand-alone statement independent of other ones; (e) the descriptor was available in English and publicly accessible. Of the materials found, a total of 282 descriptors from 13 sources met all the criteria listed above. Among them, 194 had been classified as C1 and 88 as C2. The descriptor collection consists of 5,884 lexical tokens altogether.

In order to explore the differences between the levels and then compare these differences to the common reference points induced from the ELTT scales, a keyword analysis was conducted. According to Green (2012), it is a "promising approach" (p. 94) for the purpose of identifying salient features that can potentially be useful in distinguishing one level from another. It compares the text in question with a reference corpus, generating a list of keywords that are statistically more frequent in the focus text than in the reference text. The program KeyWords Extractor, part of Cobb's (2014) Compleat Lexical Tutor available online, was used for the analysis. It "determines the defining lexis in a specialized text or corpus, by comparing the frequency of its words to the frequency of the same words in a more general reference corpus" (Cobb, 2014), in this case, the 10-million token mixed written-spoken US-UK corpus developed by Paul Nation as basis for the first 2k of the combined British National Corpus (BNC) and Corpus of Contemporary American (COCA) lists. This corpus contains 60% spoken texts including face-to-face and telephone conversations as well as movies and TV programs, and 40% written texts including fiction, letters and journals (Nation, 2018). A 'keyness factor' indicates the number of times more frequent a particular word is in the focal text than it is in the reference corpus, proportionally. For instance, if a word had 73 natural occurrences in 10,000,000 words, but 29 occurrences in a 7,613-word text, this would work out to 38,093 occurrences if the text were the same size as the corpus: (29/7,613)

Table 1. Sources of proficiency descriptors.

Source	Number of descriptors C1	C2
1. CEFR Companion Volume (new descriptors) (Council of Europe, 2018)	36	12
2. GSE Learning Objectives for Academic English	39	5
3. Profile Deutsch (translation by John Trim, as published in Green, 2012)	37	6
4. EAQUALS 2008 descriptor bank	16	18
5. ALTE Can Do Statements 2002	11	9
6. Cambridge Common Scale for Speaking	5	4
7. CEFR-J main descriptors	3	3
8. 1.2000 – Switzerland: European Language Portfolio. Version for Young People and Adults (15+)	12	6
9. ELP 84.2006 – Latvia: European Language Portfolio for Adults	7	5
10. 103.2009 – Albania: European Language Portfolio for Learners Aged 18+	7	5
11. 2014:R014 – Slovenia: European Language Portfolio 16+	9	9
12. 29.2002 – European Association of Language Centres in Higher Education (CERCLES): European Language Portfolio for University Students	8	5
13. 35.2002 – European Language Council (ELC): ELP Higher Education	4	1
Total	194	88

x 10,000,000 = 38,093; the word is thus 38,093/73 = **521.82** times more frequent in the focus text than it is in the reference corpus, which probably means that it plays an important role in the text (Cobb, 2014). The higher this figure for a given word, the more likely that word can be seen as an indicator of the "aboutness" (Baker, 2004, p. 347) of a particular text. The program routinely identifies all the words which are at least 25 times more numerous in a given text compared to the reference corpus.

Results and discussion

The analysis was conducted for C1 and C2 descriptors separately. Following Chung and Nation (2004), a more conservative approach than the program suggests was favored so as to identify the most meaningful keywords at each level; Table 2 lists all the words that have a keyness of > 50.

As can be seen, the two most important keywords that occur at both levels were *meaning* and *conversation*. The word *meaning* has a keyness factor of 32,672 and 128,571 at C1 and C2, respectively; the word *conversation* has a keyness factor of 89,848 and 85,714, respectively. That these two words are 'key' at both levels in a bank of descriptors for spoken communication is hardly surprising. Indeed, one can expect a good deal of further overlap between the two lists. Other keywords that appear at both levels include, in alphabetical order, *abstract, academic, accurate, appropriate, audience, coherent, complex, discuss, elaborate, emphasis, express, extend, flexible, fluent, grammatical, idiom, linguistic, oral, precise, professional, situation, smooth, summary* and *topic*. However, the main interest lies in the differences between the lists as the keywords unique to a particular level might have the potential to define that level more exactly. Similar to Green's (2012) analysis regarding function words, the keywords were grouped into three categories: (a) keywords unique to C1 descriptors, (b) keywords unique to C2 descriptors and (c) keywords shared between C1 and C2. Table 3 provides an overview of this categorization.

The lists in Table 3 provide some indication of significant differences between C1 and C2 in relation to speaking in academic contexts. To be able to discern the differences more easily, the keywords can be grouped according to component elements of can-do statements, including operations or activities, the object of the operation or features connected to the output text, and qualities or conditions (Green, 2012). 'Key' activities at C1 such as *paraphrase, preface, clarify, formulate* and *conclude* contrast with seemingly higher-order 'key' activities at C2, including *backtrack, rebut, pinpoint, differentiate, convey, persuade* and *eliminate*. Although it is unlikely that these operations in and of themselves are indicative of a particular level, they do reflect a notable difference in complexity. This finding is very much in line with the ELTT progression, where functions related to

Table 2. C1 and C2 keywords with a keyness factor >50, ranked in descending order.

	C1 keywords			C2 keywords		
Rank	Keyness factor	Keyword	Frequency count	Keyness factor	Keyword	Frequency count
1.	89,848.00	converse	22	128,571.00	meaning	18
2.	53,092.00	professional	13	85,714.00	converse	12
3.	32,672.00	meaning	8	21,429.00	situate	3
4.	24,504.00	relation	7	21,429.00	professional	3
5.	20,420.00	situate	5	14,285.67	idiom	6
6.	12,252.00	intone	3	3,571.50	backtrack	3
7.	6,806.67	idiom	5	2,678.62	rebut	3
8.	1,815.11	paraphrase	4	1,607.15	differentiate	9
9.	1,397.16	fluent	14	1,530.64	pinpoint	3
10.	1,225.20	nuance	3	751.87	fluent	4
11.	583.43	intelligible	4	735.29	ambiguity	8
12.	556.91	spontaneous	9	476.19	coherent	5
13.	492.05	vocabulary	10	446.44	viewpoint	3
14.	471.23	preface	3	345.63	memorable	3
15.	381.17	coherent	5	278.62	convey	11
16.	331.14	linguistic	24	227.97	grammatical	3
17.	266.35	articulate	3	219.11	shade	10
18.	222.44	complex	50	193.05	prosody	3
19.	213.82	clarify	10	174.21	elaborate	5
20.	173.79	grammatical	4	171.64	precise	13
21.	148.51	appropriate	36	164.84	oral	3
22.	142.80	abstract	10	163.40	complex	21
23.	139.62	usage	4	134.77	hostile	4
24.	133.17	seminar	3	131.73	emphasis	9
25.	127.36	formula	15	127.23	persuade	7
26.	125.66	oral	4	117.37	smooth	9
27.	114.50	flexible	9	115.91	accurate	8
28.	111.09	academy	21	106.61	eliminate	6
29.	102.10	conclude	18	101.01	appropriate	15
30.	86.43	express	44	99.90	abstract	4
31.	85.23	summary	12	96.52	linguistic	4
32.	84.97	text	16	89.01	flexible	4
33.	80.61	discourse	3	79.07	topic	9
34.	80.08	diplomat	3	65.65	confidence	5
35.	79.90	detail	27	62.66	extend	6
36.	75.49	precise	10	60.63	native	6
37.	72.93	informal	3	58.31	audience	4
38.	70.33	topic	14	54.97	express	16
39.	70.24	theme	10			
40.	66.95	discuss	34			
41.	66.68	device	4			
42.	65.68	extend	11			
43.	59.77	elaborate	5			
44.	50.01	audience	6			

The 'keyness factor' indicates the number of times more frequent a word is in the focal text than it is in the reference corpus.

linguistic planning and repair are clearly related to the two lower bands, whereas more nuanced functions such as persuasion and differentiation are associated with the highest bands. Keywords that can be related to the object of the operation or output text, in contrast, do not seem to advance our understanding of the level characteristics in any significant way. At C1, such words include *intonation, vocabulary, text* and *discourse*; at C2 only *viewpoint* and *prosody* occurred, possibly alluding to the picture that emerged in the ELTT scales where the range of prosodic features increases as we move up the scale. Keywords referring to qualities or conditions have perhaps the greatest potential to characterize the levels more fully. *Relation, nuance, intelligible, detail, usage, spontaneous, diplomat* and *informal* at C1 contrast with *ambiguity, memorable, shade, hostile, native* and *confidence* at C2.

Since a keyword analysis focuses only on lexical rather than semantic differences (Baker, 2004), the true meaning of the keywords and the potential for them to be criterial can, of course, only be

Table 3. Keywords unique to a particular level or shared by both.

C1	C1 and C2	C2
relation	abstract	backtrack
intone	academy	rebut
paraphrase	accurate	differentiate
nuance	appropriate	pinpoint
intelligible	audience	ambiguity
spontaneous	coherent	viewpoint
vocabulary	complex	memorable
preface	converse	convey
articulate	discuss	shade
clarify	elaborate	prosody
usage	emphasis	hostile
seminar	express	persuade
formula	extend	eliminate
conclude	flexible	confidence
text	fluent	native
discourse	grammatical	
diplomat	idiom	
detail	linguistic	
informal	meaning	
theme	oral	
device	precise	
	professional	
	situate	
	smooth	
	summary	
	topic	

ascertained by considering the co-text as well. Therefore, the descriptor text was transformed into a complete concordance index for every word, using Text-Based Concordances available on the Compleat Lexical Tutor website (Cobb, 2016). A qualitative observation of all the concordance lines for the C2 keyword *ambiguity*, for example, shows that it invariably collocates with *eliminate* or *without*. Accordingly, the ability of a learner to resolve alternative, even competing interpretations in real time or to reduce the vagueness inherent in spoken communication might be considered a characteristic feature at level C2. This result is hardly surprising given the frequency and prominence of this collocation in the original CEFR descriptors. It can, however, offer some insight as to how important and stable the concept has become at that level. Similarly, the keyword *shade* always appears in the phrase *can convey finer* ***shades*** *of meaning precisely*, which consolidates its position as a characteristic feature at level C2. An example of an extra dimension emerging from the analysis is the keyword *confidence*. In the original CEFR descriptors, the idea of confidence features only once at level C2: *can present a complex topic* ***confidently*** *and articulately to an audience unfamiliar with it*, an illustrative descriptor for addressing audiences. In the present analysis, however, *confidence* is a keyword, occurring repeatedly in different contexts. This contrasts markedly with C1 descriptors, where *confidence* occurs only once as the object of an operation in connection with expressing degrees of confidence or uncertainty, but never referring to the quality of the performance.

For the keywords shared between C1 and C2 descriptors, listed in the middle column of Table 3, there would seem to be at least four possible explanations: firstly, such words are 'key' in the descriptor formulations but do not represent a salient performance feature; secondly, such keywords represent salient features at both levels; thirdly, they represent a transition stage between C1 and C2; or fourthly, their classification may ultimately depend on the collocates that go with them. Again, a concordance analysis can shed some light on the matter. As one would expect, words like *academic, discuss, grammatical, linguistic* and *oral* play an important role at both levels without adding anything of substance to the level specifications. The word *situation* may potentially be significant as it could relate to the contexts or conditions in which the communication is embedded,

and these could be different at the lower vis-à-vis the higher level. However, the concordance table did not reveal a consistent pattern in this respect. Words related to the categories of communicative language competences in the CEFR do not show a distinct pattern either; they all seem to be important at both levels, although with slightly different keyness values at C1 and C2, respectively: *fluent* (1,397.16– 751.87), *coherent* (381.17– 476.19), *complex* (222.44– 163.40), *abstract* (142.80– 99.90), *appropriate* (148.51– 101.01), *accurate* (41.42– 115.91). The somewhat more marked differences in keyness of the words *fluent* and *accurate* could indicate that fluency in a second language reaches a natural plateau at C1 and that disfluencies at C2 are those characteristic of spoken discourse also observable in highly proficient speakers, whereas C2 speech is characterized by an ever-increasing degree of accuracy. The latter would seem to be in line with North's (2014) conclusion drawn from the evidence he cites that it is mainly a surge in accuracy that distinguishes the C levels rather than the mastery of new features. This interpretation is speculative, of course, and whether these words are (part of) qualifiers representing a transition level or whether they are equally important at both levels cannot be answered definitively from this analysis.

The remaining keywords, *idiom, precise, flexible* and *audience*, however, do show a pattern. While at C1, *idiomatic* is always accompanied by a qualification (e.g. *a variety of* ***common*** *idiomatic expressions; can* ***in most cases*** *understand idiomatic expressions; idiomatic expressions* ***in my field***), at C2 such qualifications are entirely absent. On the contrary, the surrounding text has an intensifying function (e.g. *a* ***good command*** *of idiomatic expressions; a* ***good familiarity*** *with idiomatic expressions; even when the debate is* ***highly*** *idiomatic*). The keyword *precise* occurs in various contexts at level C1; at C2, it is almost invariably part of the phrase *convey finer shades of meaning* ***precisely***. The keyword *flexible* repeatedly collocates with the emphasizer *very* at level C2, but not so at C1. Finally, the keyword *audience* at C1 is connected to the ability to respond to questions from or points raised by the audience and to structure speech in a listener-friendly way, whereas the C2 descriptors are exclusively about the ability to adapt to the specific needs of the audience (e.g. *when it is clear that the* ***audience*** *needs it; adapting the talk flexibly to meet the* ***audience's*** *needs; tailoring my presentation to the* ***audience****; to an* ***audience*** *unfamiliar with it*).

The relationship between the keywords and the common reference points

The differences between C1 and C2 as revealed through the keyword analysis seem to tally with the common reference points generated in the scale validation project. Although there is no one-to-one correspondence between the common reference points for five bands and the three categories of keywords presented in Table 3, it is fair to assume that C1 keywords can be compared to the two lowest ELTT bands and C2 keywords to the two highest ones. From this perspective, the level-specific keywords are at least indicative of the proposed progression. They appear to reflect the overall tendency that, as proficiency increases, the focus shifts from linguistic aspects like accuracy and appropriateness to communicative impact and effectiveness. The common reference points in the bands termed 'operational proficiency' largely pertain to linguistic appropriateness, errors, planning and repair, which is mirrored by the overall extent to which the keywords at C1 denote linguistic concepts: *intonate, paraphrase, vocabulary, articulate, usage, formulaic, text, discourse, informal*. This is in line with the salient features of spoken language at C1 as described in the CEFR (Council of Europe, 2001, p. 36, 2018, pp. 150–151): "good access to a broad range of language, which allows fluent, spontaneous communication" and discourse skills "with an emphasis on more fluency", as illustrated by descriptor elements like "can express him/herself fluently and spontaneously, almost effortlessly", "gaps to be readily overcome with circumlocutions" or "little obvious searching for expressions or avoidance strategies".

The common reference points in the bands termed 'academic proficiency', in contrast, relate predominantly to the skill in using language and the communicative effect, notably consistent control, automaticity, ease, flexibility as well as full/deliberate/skillful/effective use of communicative means, which is mirrored in the C2 keywords by the extent to which they indicate functional

Bands		Common reference points
1	Full Academic	• high degree of automaticity and flexibility in language use • consistent phonetic/phonological control • deliberate use of vocal features • full use of presentation skills • flexibility and ease even with abstract, complex unfamiliar topics • skillful and effective use of sophisticated collaboration strategies
2	Advanced Academic	• consistent accuracy even when the focus is not on language • fluency, flexibility and ease • meta-cognitive awareness to enhance communicative effect • initiative and interaction management • effective persuasion • effective use of collaboration strategies
3	General Academic	• disfluencies merely for communicative enhancement • effective argumentation, logic and valid reasoning • attributes of professional public speakers begin to emerge • adaptive responsiveness • use of advanced collaboration strategies
4	Full Operational	• few disfluencies for linguistic planning and repair • errors are insignificant • appropriate use of basic task-specific activities • discernible awareness of contextual factors • use of basic collaboration strategies
5	Effective Operational	• clarity and appropriateness of language • disfluencies for linguistic planning and repair • appropriate use of essential academic functions

Figure 1. Common reference points for academic speaking at C1–C2 (based on Berger, 2015, p. 289).

subtleties: *backtrack, rebut, differentiate, pinpoint, persuade, eliminate, ambiguity, memorable, shade*. Again, this is similar to the salient features of spoken language as described in the CEFR (Council of Europe, 2001, p. 36, 2018, p. 151): at C2 they include "the degree of precision, appropriateness and ease" typical of "highly successful learners", as illustrated by descriptors like "convey finer shades of meaning precisely" or "a good command of idiomatic expressions and colloquialisms with awareness of connotative level of meaning".

Besides this general pattern, there are also some very close correspondences between the keywords and the common reference points. The C1 keyword *clarify*, for example, alludes to the clarity of the language characteristic of *Effective Operational Proficiency*. The keyword *paraphrase* seems to correspond to the linguistic planning and repair mentioned in the two lowest ELTT bands. At the other end of the spectrum, the C2 keyword *persuade* has a direct counterpart in the *Advanced Academic Proficiency* band, namely effective persuasion. The coherence between the keywords and the common reference points is thus very high, with possibly one discrepancy. The C1 keyword *nuance* is closely related to other C2 keywords (*differentiate, shade*), matching the common reference points in the higher bands better than those in the lower bands. Integrating salient keywords with the common reference points (as presented in Figure 1), the progression can be summarized as follows (based on Berger, 2015):

In the lowest band, termed ***Effective Operational Proficiency***, learners can speak with sufficient control to participate productively in formal university settings. Their linguistic repertoire allows them to express conceptually complex ideas clearly and appropriately, with occasional disfluencies owing to linguistic planning and repair. They can use essential communicative functions in academic presentations and interactions.

The next higher band, ***Full Operational Proficiency***, reflects the learners' ability to speak entirely appropriately in formal university settings. The progression is characterized by an increased control and sophistication of the linguistic resources. Learners can use an even larger range of language allowing them to decrease the number of errors as well as pauses due to linguistic planning and repair, with errors being insignificant. They can also use a greater variety of task-related functions. What is new in this band is a discernible awareness of contextual factors such as audience and task requirements.

In the next band, ***General Academic Proficiency***, learners can speak effectively for most academic purposes. It marks a new quality dimension both linguistically and strategically. Disfluencies mainly occur due to linguistic refinement rather than planning and repair. In addition, more complex cognitive skills integral to academic practice start to emerge in this band: there is a new focus on argumentation, aspects of logic and reasoning. The speakers' attentional focus seems to shift from communicative appropriateness to strategic effectiveness. Attributes of professional public speakers begin to feature in this band. Finally, in relation to interaction skills, learners demonstrate a new degree of adaptive responsiveness; that is, they can react to unexpected circumstances by picking up contextual cues and adapting their contributions accordingly.

In the next band, ***Advanced Academic Proficiency***, learners can speak fluently and accurately on all levels pertinent to academic presentations and interactions. They can use a very broad range of language, which allows them to speak without having to restrict the effectiveness of their contribution. They can speak naturally, fluently and with a consistently high degree of accuracy, even when their cognitive resources are being directed towards non-linguistic matters. There are two new features: Firstly, there is an unprecedented degree of flexibility and ease, and secondly, the descriptors reflect a new degree of meta-cognitive awareness that allows speakers to perform their tasks more consciously and effectively. In functional terms, this band is characterized by an increased capacity for persuasion rather than just argumentation and for interaction management rather than just participation.

The highest band, ***Full Academic Proficiency***, represents the highest degree of control, precision, flexibility, naturalness and confidence in all respects. Learners can use a very broad range of language that is consistently accurate and highly idiomatic. The speaking process is completely automatic so that the learners can direct their full attention to the communicative effect.

Conclusion

This article has sketched out the path from context-specific, assessor-oriented rating scale descriptors for academic presentations and interactions in English as a foreign language to a more context-neutral, constructor-oriented set of common reference points characterizing academic speaking proficiency in abstraction from the specifics of the local context. It has presented the findings of a keyword analysis of C1 and C2 descriptors for speaking taken from several CEFR-based proficiency scales relevant to learners of English in adult or higher education. The main aim of the study was to validate the locally-developed progression of speaking ability as represented in the common reference points that had been extracted from an existing set of calibrated rating scale descriptors.

As it turns out, the keywords approach is a relatively quick and easy method of revealing central concepts in a given collection of texts in relation to other texts. As such, it is useful for the purpose of identifying salient characteristics of level descriptors. Indeed, the analysis indicates that a number of keywords in the descriptor bank are related to the common reference points. Some keywords clearly overlap with salient characteristics of the rating scale bands; others are roughly related. These similarities suggest the soundness of the progression in the ELTT scales in the most general terms.

The comparison between the keyword lists and the common reference points can be no more than indicative, though. Keywords from C1 and C2 descriptors without any further subdivisions cannot be directly related to salient characteristics of a five-band progression. In addition, there are conceptual and methodological limitations that warrant caution when interpreting the results. For

the present purposes, the basic principle of keywords turns out to be a disadvantage: keywords are based on lexical but not semantic or functional differences. Consequently, synonymous concepts will not appear to be key although they may potentially be characteristic of a particular level. The function of leading a group discussion is a case in point. Several C2 but no C1 descriptors refer to active interaction management, albeit in different formulations, e.g. *can lead a panel discussion on abstract or academic topics; can steer conversation in a larger group; can act as moderator; can take on different roles according to the needs of the participants and requirements of the activity (resource person, mediator, supervisor, etc.); can recognize undercurrents in interaction and take appropriate steps accordingly; can guide the direction of the talk; can effectively lead the development of ideas in a discussion*. Level characteristics based on keywords may therefore overattend to lexical differences while at the same time overlook key conceptual features.

Another weakness is the narrow basis of the keyword analysis. Being limited to proficiency scales from publicly available documents, the analysis is based on a possibly non-representative snapshot of the 'universe' of proficiency descriptions. Additional can-do statements from a larger number and wider range of sources may well have resulted in different keywords or keyness factors. Furthermore, the potential that collective and established knowledge available through materials designed for pedagogic and assessment purposes, such as course syllabi, textbooks and test specifications, may have for present purposes was not tapped. A more comprehensive research design that synthesizes the frequency-based definition of saliency underlying the keywords approach with alternative concepts of saliency, based on expert judgment for example, would lead to a fuller description of how speaking proficiency advances. Including other multi-level assessment scales, in particular, would allow a more direct comparison with the common reference points. Finally, using a different statistical procedure for calculating keyness, such as Dunning's Log Likelihood function or chi-square (Scott & Tribble, 2006), as well as a different reference corpus that is more closely related to the target corpus (Scott, 2009) would likely diminish some of the methodological limitations.

What the limited source base implies is a certain circularity in the approach. Where proficiency scales linked to the CEFR are used to validate common reference points induced from specific descriptors representing CEFR levels, the approach seems to contain an assumption of what is to be investigated. Green (2012) reminds us that some circularity is inherent in such processes of refinement. Indeed, the approach can be considered circular to the extent that the source and reference texts are overly similar and interrelated, and that the descriptors represent a duplication rather than an agreed interpretation of the CEFR levels. However, although many of the can-do statements included in the descriptor bank are derived from or informed by the CEFR, most schemes have modified, expanded and elaborated the CEFR descriptors, adding entirely new dimensions to meet the specific needs of the target group. This flexibility and openness, which is a basic principle of the CEFR (Council of Europe, 2001), may ensure that the level interpretations in the sources are sufficiently varied and eventually help to generate a more comprehensive representation of currently operational levels. Where there is repetition and overlap with CEFR formulations in the descriptor bank, this may indicate the importance of the concepts concerned, and identifying these concepts may enhance our understanding of which aspects are robust and which are less clear or unequivocal. From this perspective, the purpose of the keyword analysis was not only to validate the common reference points but also to confirm and possibly enrich our understanding of the levels that the CEFR promotes.

A fuller understanding of the higher levels is crucial for learners and educators alike. Generic reference points such as those derived from the specific descriptors can serve as benchmark statements for future construct definition and scale development projects. They also signpost the path university students could follow if they wanted to improve their speaking proficiency, and teachers can organize their lessons accordingly. Such specification addresses the urgent need in higher education for empirically based instruments that are sensitive enough to capture also slight gains in learning. While this study has provided some preliminary insights into key concepts in proficiency descriptors for speaking, thereby reflecting, at least in general terms, the overall soundness of the ELTT progression, the keywords approach is necessarily tentative and must be complementary to other forms of validation.

Disclosure statement

No potential conflict of interest was reported by the author.

References

Adey, P. (1997). Dimensions of progression in a curriculum. *The Curriculum Journal*, *8*(3), 367–391. doi:10.1080/0958517970080304

Alderson, C. (1991). Bands and scores. In C. Alderson & B. North (Eds.), *Language testing in the 1990s: The communicative legacy* (pp. 71–94). London, UK: Macmillan Publishers.

Archer, D., Culpeper, J., & Rayson, P. (2009). Love – 'a familiar or a devil'? An exploration of key domains in Shakespeare's comedies and tragedies. In D. Archer (Ed.), *What's in a word-list? Investigating word frequency and keyword extraction* (pp. 137–157). Farnham, UK: Ashgate.

Bailey, A., & Heritage, M. (2014). The role of language learning progressions in improved instruction and assessment of English language learners. *TESOL Quarterly*, *48*(3), 480–506. doi:10.1002/tesq.176

Baker, P. (2004). Querying keywords: Questions of difference, frequency, and sense in keywords analysis. *Journal of English Linguistics*, *32*(4), 346–359. doi:10.1177/0075424204269894

Baker, P. (2009). 'The question is, how cruel is it?' Keywords, fox hunting and the House of Commons. In D. Archer (Ed.), *What's in a word-list? Investigating word frequency and keyword extraction* (pp. 125–136). Farnham, UK: Ashgate.

Berger, A. (2015). *Validating analytic rating scales: A multi-method approach to scaling descriptors for assessing academic speaking*. Frankfurt am Main, Germany: Peter Lang.

Berger, A. (2018). Rating scale validation for the assessment of spoken English at tertiary level. In G. Sigott (Ed.), *Language testing in Austria: Taking stock* (pp. 679–702). Frankfurt am Main, Germany: Peter Lang.

Berger, A., & Heaney, H. (2018). Developing rating instruments for the assessment of academic writing and speaking at Austrian university English departments. In G. Sigott (Ed.), *Language testing in Austria: Taking stock* (pp. 325–346). Frankfurt am Main, Germany: Peter Lang.

Capel, A. (2015). The English vocabulary profile. In J. Harrison & F. Barker (Eds.), *English profile in practice* (pp. 9–27). Cambridge, UK: Cambridge University Press.

Chung, T. M., & Nation, P. (2004). Identifying technical vocabulary. *System*, *32*(2), 251–263. doi:10.1016/j.system.2003.11.008

Cobb, T. (2014). *KeyWords extractor* (Version 2) [Software]. Retrieved from https://www.lextutor.ca/key/

Cobb, T. (2016). *Text-Based concordances* (Version 3) [Software]. Retrieved from https://www.lextutor.ca/conc/text/

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

Council of Europe. (2018). *Common European framework of reference for languages: Learning, teaching, assessment: Companion volume with new descriptors*. Retrieved from https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

Deygers, B., Zeidler, B., Vilcu, D., & Carlsen, C. H. (2017). One framework to unite them all? Use of the CEFR in European university entrance policies. *Language Assessment Quarterly*, *15*(1), 3–15. doi:10.1080/15434303.2016.1261350

Fulcher, G. (2012). Scoring performance tests. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 378–392). London, UK: Routledge.

Fulcher, G. (2016). Standards and frameworks. In J. Banerjee & D. Tsagari (Eds.), *Handbook of second language assessment* (pp. 29–44). Berlin, Germany: DeGruyter Mouton.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, *28*(1), 5–29. doi:10.1177/0265532209359514

Green, A. (2012). *Language functions revisited: Theoretical and empirical bases for language construct definition across the ability range*. Cambridge, UK: Cambridge University Press.

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

Hulstijn, J. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal*, *91*(4), 663–667. doi:10.1111/modl.2007.91.issue-4

Hulstijn, J. H., Alderson, C., & Schoonen, R. (2010). Developmental stages in second-language acquisition and levels of second-language proficiency: Are there links between them? In I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency in linguistic development: Intersections between SLA and language testing research* (Eurosla Monographs Series 1) (pp. 11–12). European Second Language Association. Retrieved from http://eurosla.org/monographs/EM01/11-20Hulstijjn_et_al.pdf

Jones, N., & Saville, N. (2008). Scales and frameworks. In B. Spolsky & F. Hult (Eds.), *The handbook of educational linguistics* (pp. 495–509). Oxford, UK: Blackwell Publishing.

Kaftandjieva, F., & Takala, S. (2002). Council of Europe scales of language proficiency: A validation study. In C. Alderson (Ed.), *Common European framework of reference for languages: Learning, teaching, assessment: Case studies* (pp. 106–129). Strasbourg, France: Council of Europe.

Linacre, J. M. (2013). *Facets computer program for many-facet Rasch measurement* (Version 3.71.2.) [Software]. Retrieved from http://www.winsteps.com

Luoma, S. (2004). *Assessing speaking*. Cambridge, UK: Cambridge University Press.

Milanovic, M. (2009). Cambridge ESOL and the CEFR. *Research Notes*, *37*, 2–5. Retrieved from http://www.cambridgeenglish.org/images/23156-research-notes-37.pdf

Nation, P. (2018). *The BNC/COCA word family lists*. Retrieved from https://www.victoria.ac.nz/__data/assets/pdf_file/0004/1689349/Information-on-the-BNC_COCA-word-family-lists-20180705.pdf

North, B. (2000). *The development of a common framework scale of language proficiency*. New York, NY: Peter Lang.

North, B. (2014). *The CEFR in practice*. Cambridge, UK: Cambridge University Press.

North, B., Ortega, A., & Sheehan, S. (2010). *British Council – EAQUALS core inventory for general English*. London: British Council/EAQUALS. Retrieved from http://englishagenda.britishcouncil.org/sites/default/files/attachments/books-british-council-eaquals-core-inventory.pdf

O'Keeffe, A., & Mark, G. (2017). The English grammar profile of learner competence: Methodology and key findings. *International Journal of Corpus Linguistics*, *22*(4), 457–489. doi:10.1075/ijcl.14086.oke

O'Sullivan, B. (2014). Assessing speaking. In A. Kunnan (Ed.), *The companion to language assessment* (pp. 156–171). Chichester, UK: Wiley Blackwell.

Saville, N. (2012). The CEFR: An evolving framework of reference. In E. Tschirner (Ed.), *Aligning frameworks of reference in language testing: The ACTFL proficiency guidelines and the common European framework of reference for languages* (pp. 57–69). Tübingen, Germany: Stauffenburg Verlag.

Scott, M. (2009). In search of a bad reference corpus. In D. Archer (Ed.), *What's in a word-list? Investigating word frequency and keyword extraction* (pp. 79–91). Farnham, UK: Ashgate.

Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education*. Amsterdam, Netherlands: John Benjamins.