Article

LANGUAGE
TESTING

Language Testing
2017, Vol. 34(4) 477–492
© The Author(s) 2017
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0265532217711431
journals.sagepub.com/home/ltj
SAGE

Language assessment and the inseparability of lexis and grammar: Focus on the construct of speaking

Ute Römer
Georgia State University, USA

Abstract

This paper aims to connect recent corpus research on phraseology with current language testing practice. It discusses how corpora and corpus-analytic techniques can illuminate central aspects of speech and help in conceptualizing the notion of lexicogrammar in second language speaking assessment. The description of speech and some of its core features is based on the 1.8-million-word Michigan Corpus of Academic Spoken English (MICASE) and on the 10-million-word spoken component of the British National Corpus (BNC). Analyses of word frequency and keyword lists are followed by an automatic extraction of different types of phraseological items that are particularly common in speech and serve important communicative functions. These corpus explorations provide evidence for the strong interconnectedness of lexical items and grammatical structures in natural language. Based on the assumption that the existence of lexicogrammatical patterns is of relevance for constructs of speaking tests, the paper then reviews rubrics of popular high-stakes speaking tests and critically discusses how far these rubrics capture the central aspects of spoken language identified in the corpus analyses as well as the centrality of phraseology in language. It closes with recommendations for speaking assessment in the light of this characterization of real-world spoken lexicogrammar.

Keywords

Assessment rubrics, corpus analysis, lexicogrammar, phraseology, speaking assessment, spoken corpora

Historically, lexis and grammar have been treated separately in language testing. This separation can be seen both in models of language ability that are used to inform test development and validation, and in many frequently used rating scales that are used to

Corresponding author:
Ute Römer, Georgia State University, Department of Applied Linguistics and ESL, 25 Park Place, Suite 1500, Atlanta, GA 30303, USA.
Email: uroemer@gsu.edu

score performance assessments. The reasons for this separation likely lie in the history of language testing researchers' understanding of language proficiency, in other words their answer to the question "What does it mean to know a language?" Coming from a structuralist tradition within linguistics (Bloomfield, 1933), early language testing researchers such as Lado (1961) promoted a skills/components model of language, in which grammatical structure is conceptually distinct from vocabulary, and recommended testing these components separately. More recent models of language ability, including the influential model of Bachman and Palmer (1996, 2010), continue this separation of lexis and syntax as distinct aspects of "grammatical knowledge", separating these aspects of language ability from knowledge of language functions, which is subsumed under "pragmatic knowledge".

Based on this view of language, many influential rating scales in language testing have traditionally treated lexis and grammar separately. An early rating scale, the Interagency Language Roundtable (ITR) scale, developed by the US Foreign Service Institute in the 1950s to assess a speaker's spoken language proficiency, refers to errors of grammar independently from breadth and precision of vocabulary. Other speaking tests developed in the second half of the 20th century also appear to support a view of lexicogrammar that is binary and considers the correctness of grammatical structures and vocabulary separately (see, e.g., Luoma, 2004; Taylor, 2011; Vidaković & Galaczi, 2013).

Within more recent integrative and functionally oriented approaches to language learning, scholars have begun to look at language proficiency more holistically and consider lexicogrammatical knowledge (capturing forms and functions of words, phrases and utterances) as a single category. Corpus linguists, including Sinclair (2008), argue that the phrase, rather than the individual word, is the fundamental unit of language, and that a great deal of communication consists of memorized fixed expressions that defy simple categorization into either lexis or grammar (see also Biber, 2009; Ellis, Römer, & O'Donnell, 2016; Hoey, 2005; Hunston, 2002; Hunston & Francis, 2000; Römer, 2005, 2009, 2010; Sinclair, 1991, 2004; Stubbs, 2001).

This interdependence of grammar and lexis is only beginning to be recognized within some areas of language testing. For example, Alderson and Kremmel (2013) question the usefulness of separating vocabulary and grammar in the context of testing reading ability and raise the concern that lexicogrammar "should perhaps instead be treated as a unitary component of reading ability rather than attempting to distinguish between vocabulary and grammar" (p. 550; see also Shiotsu & Weir, 2007). However, within speaking assessment, it is not clear that this more holistic view of lexicogrammar has taken hold, particularly with regard to rating scale development. According to Fulcher (2003, p. 14), a major problem with many rating scales is that their descriptors are not based on analyses of empirical linguistic evidence but "come from intuitive judgements about how language competency develops". Corpus studies of lexicogrammar may provide such empirical evidence that may be useful in informing the development, validation, and use of rating scales for speaking assessment.

With this issue in mind, the present article discusses the importance of considering corpus evidence in highlighting central aspects of spoken language in the context of speaking assessment. This paper has two goals. First, I demonstrate how corpus tools and techniques can be used to investigate patterns of language use that cannot be neatly separated into lexis and grammar, and highlight the pervasiveness of such patterns in oral language. Second, I address the question: "Do rating scales of current speaking tests capture the core features of lexicogrammar as identified in corpus analyses?" As the title of this article suggests, the focus is solely on spoken English lexis and grammar (considered as a unit), and not on other factors that affect the ratings of learner speech, such as pronunciation, speech rate, intonation, or intelligibility.

This paper examines the role of lexicogrammar in spoken English through the study of two of the largest freely available English speech corpora (further described below): the Michigan Corpus of Academic Spoken English (MICASE; Simpson, Briggs, Ovens, & Swales, 2002) and the spoken component of the British National Corpus (BNC_spoken; Burnard, 2007). These corpora contain samples of language from a variety of academic and general purpose speech events, respectively, and were therefore selected to be representative of two major domains of language use associated with the two purposes for which oral language is typically assessed: for admissions and placement in academic settings and to obtain scores for general oral proficiency. To assess the representation of lexicogrammatical knowledge in current speaking tests, the article then reviews rating scales of a selection of high-stakes speaking tests developed by three major international testing companies, including the TOEFL iBT (Educational Testing Service, ETS), the ECCE and ECPE (Cambridge Michigan Language Assessments, CaMLA), the IELTS, and the Cambridge English: Advanced (CAE) exam (both Cambridge English Language Assessment). It examines in how far these rating scales capture central aspects of spoken English as highlighted by the MICASE and BNC_spoken explorations. This paper closes with a discussion of implications of our corpus-based findings for speaking assessment and a brief overview of required future research on the topic.

The construct of speaking from a corpus perspective

The goal of the corpus analyses described in the following sections is to illuminate the construct of speaking and identify central aspects of authentic spoken lexicogrammar across a wide variety of speech events in both general and academic English. The core corpus-analytic techniques used to achieve this goal include creating and examining frequency word and keyword lists, and extracting phraseological items (n-grams and phrase-frames) from the corpora. The resulting lists are then manually analyzed in order to identify items that play a central role in speech and carry important discourse functions.

Corpora and tools for analysis

Two types of corpora were used to conduct the analyses: corpora containing samples of spoken language and corpora of written language to serve as a point of reference for comparison. The two corpora of spoken language selected for this study are the Michigan Corpus of Academic Spoken English (MICASE) and the spoken section of the British National Corpus (BNC_spoken). MICASE is a collection of 152 transcripts and 1.8 million words of academic speech, based on 200 hours of recordings of speech events from across a US research university, the University of Michigan in Ann Arbor. The corpus captures authentic spoken discourse from a range of academic settings (e.g., lectures, seminars, and study groups) and academic disciplines distributed fairly evenly across the humanities, social sciences, health sciences, and physical sciences. BNC_spoken consists of 915 transcripts and 10 million words of general British English speech from a variety of contexts and speakers across the British Isles. It contains spoken material from informal, spontaneous conversations and from more formal events including interviews, business meetings, and lectures.

Two corpora of written English were chosen as reference corpora to serve in the creation of the keyword lists. The keyword analysis will therefore be able to identify features of lexicogrammar that distinguish speech from writing. Chosen to serve as reference were analogous corpora of written English: the Hyland Corpus of academic writing (Hyland, 1998) used as a reference corpus for MICASE, and the written component of the British National Corpus (BNC_written) used as a reference corpus for BNC_spoken.

The software tools selected for corpora access and analysis are *AntConc* (version 3.4.3; Anthony, 2014) and *kfNgram* (Fletcher, 2007), as described below.

Word frequency and keyword lists

The first step taken in approaching MICASE and BNC_spoken was the generation of a frequency word list for each corpus. Such lists are useful in the identification of aspects of a specific domain of language use because they highlight which words are most common in a corpus and possibly serve important functions in the discourse that the corpus represents. The frequency word lists for both BNC_spoken and MICASE show that the most common words in academic and general spoken English are essentially "small words", including articles (*a, the*), prepositions (e.g., *in, of*), conjunctions (*and, but*), and pronouns (e.g., *I, you, we*). Fifteen out of the 20 most frequent items are shared across the two corpora. These are words that are generally frequent in English, and most of them would also appear near the top of a word list based on a corpus that contains written material. Exceptions may be the items *so* and *like* (MICASE) and the hesitation markers *uh* (MICASE) and *er* (BNC_spoken). These are likely to be words that are characteristic of speaking rather than writing. This hypothesis can be confirmed in the next analytic step: the creation of keyword lists.

Although a word list highlights what is frequent in a corpus (and hence in a particular text type or register), it does not necessarily show what is important or unusually frequent. In other words, it does not indicate which items are characteristic of the type of language the corpus aims to capture. One way to identify those items is to compare a frequency word list based on a target corpus (here MICUSP and BNC_spoken) with another frequency word list based on a reference corpus and thus create a keyword list (e.g., Scott & Tribble, 2006). In a keyword list, such as the one displayed in Table 1, words are usually listed in order of their keyness values. Words get a high keyness value if they occur considerably more frequently in a selected corpus than they would be expected to occur on the basis of figures derived from a reference corpus.

The two lists in Table 1 show the 20 most "key" words in MICASE and BNC_spoken, with the Hyland Corpus and BNC_written used as reference corpora. All of the listed

Table 1. The top-20 keywords in MICASE (RC: Hyland Corpus) and BNC_spoken (RC: BNC_written), ordered by keyness value.

Rank	MICASE			BNC_spoken		
	Keyword	Frequency	Keyness	Keyword	Frequency	Keyness
1	you	37,835	35,465.1	er	73,656	326,840.6
2	I	33,840	22,480.4	you	208,921	315,650.9
3	it's	12,605	12,730.2	I	239,113	309,642.3
4	uh	11,277	12,294.7	erm	50,115	224,550.6
5	so	16,694	11,499.5	yeah.	39,353	178,391.1
6	um,[1]	10,410	11,465.5	oh	41,226	170,586.2
7	like	11,432	8,276.3	it's	66,991	157,333.3
8	know	9419	8234.3	that's	44,450	132,492.4
9	that's	7818	8086.6	yeah	25,633	116,026.0
10	yeah	6956	7661.4	got	46,719	102,144.7
11	um	6760	7162.3	well	51,376	98,332.2
12	okay	5531	6022.8	mm.	20,795	92,153.9
13	what	11,950	5870.4	don't	40,630	72,716.8
14	(xx)	5299	5752.1	cos	15,737	68,271.3
15	just	8592	5733.2	think	37,493	66,905.8
16	don't	6432	5716.3	er,	14,529	64,263.0
17	uh,	5170	5686.9	we	76,174	61,412.6
18	think	6192	4840.6	yeah,	13,817	60,370.3
19	I'm	4571	4371.9	know	36,916	59,596.9
20	gonna	4029	4365.8	yes.	14,918	58,437.2

words are comparatively much more frequent in (academic) spoken than in written English. They hence help to highlight aspects that are characteristic of speech as opposed to writing. Topping both keyword lists are the personal pronouns *you* and *I*, which are considerably more common in the type of language captured in MICASE and BNC_spoken than in written corpora. As Leech (2000, p. 694) points out, "conversational grammar reflects a shared context", which explains the frequent use of deictic markers and pronouns. Also very high up in both keyness-sorted lists are the hesitation markers *uh* and *um* (MICASE), and *er* and *erm* (BNC_spoken), with the different spellings being due to different transcription conventions followed in the compilation of the two corpora. These hesitation markers are often followed by a comma (e.g., *uh*, in MICASE) representing a brief mid-utterance pause. Hesitation is part of fluent native-speaker discourse. The use of markers such as *erm* and *uh* give speakers planning time and allow them to organize their thoughts (see also Fulcher, 1996). Other items in the lists include short forms such as *it's, don't, cos, I'm*, and *gonna*, and the response tokens or backchanneling devices *yeah, okay, mm*, and *yes*. The high frequency of those items in the spoken corpora provides support for Leech's observations that "conversational grammar is interactive grammar" (2000, p. 696) and that it is "adapted to the needs of real-time processing" (p. 698). Also key in spoken English are discourse markers (*so, like, well*), often used as cohesive devices, and laughter, transcribed as (xx) in MICASE. The remaining items in the lists include *what*, and the verb forms *know, think*, and *got*. An inspection of this list suggests that these words are particularly frequent in speech because they are part of frequently used phrases that are typical of speaking as opposed to writing; confirming this suspicion requires an additional step, that is, the analysis of phrases rather than individual words.

The items in the MICASE and BNC_spoken word and keyword lists clearly point to some core register differences between speech and writing, confirming earlier empirical work on the topic (Biber, 1988; Biber, Leech, Johansson, Conrad, & Finegan, 1999; Carter & McCarthy, 1995, 2006; Leech, 2000). They do, however, not provide any major insights into spoken discourse functions and into how meanings are expressed in speech. To address this issue, I will now turn to extracting larger linguistic units (i.e., different types of phraseological items) from the two corpora.

Phraseological items: n-grams and phrase-frames

The corpus analysis so far has focused only on frequent words and keywords in isolation. However, spoken communication does not usually consist of single words but combinations of those. Also, as Sinclair (2008, p. 409) pointed out, "the normal primary carrier of meaning [in language] is the phrase and not the word; the word is the limiting case of the phrase, and has no other status in the description of meaning." Hence, to investigate the creation of meaning in the discourse and in communicative functions of language, it is necessary to look beyond the word and at larger units. The next corpus-analytic steps therefore focus on phraseological items (variably referred to as n-grams, formulaic sequences, collocational frameworks, lexical bundles, clusters, etc.) that are particularly common in speaking. Two specific phraseological items that can be automatically extracted from texts are n-grams, which are contiguous word sequences of different lengths (e.g., *you know, a lot of*), and phrase-frames, which are non-contiguous word sequences (e.g., *a * of, I don't * if*), The programs *AntConc* and *kfNgram* were used to extract n-grams and phrase-frames of different lengths from MICASE and BNC_spoken.

Examined were frequency-sorted lists of 2-grams, 3-grams, and 4-grams (i.e., repeated sequences of two, three, and four words) created with the help of the "Clusters/N-Grams" function in *AntConc*. Examples of 2-grams that are common in both corpora include *of the, in the, you know, I think, I don't, I mean*, and *you can*. The 20 most frequent 3-grams in MICASE and BNC_spoken are listed in Table 2. Items that are shared across both lists include *a lot of, I don't know, one of the, you have to, I don't think*, and *I think that*. Other items are very frequent in the academic speech corpus but less so in BNC_spoken, for example *this is a, in terms of, you know what, part of the, you can see*, and *you need to*. A large number of the most frequent 4-grams in MICASE and BNC_spoken are extensions of items from the 3-gram lists displayed in Table 2, for example *I don't know if, I don't know how, a lot of the, the end of the, at the end of, you know what I, one of the things*, and *to be able to*. Other frequent spoken 4-grams are *at the same time, if you look at, have a look at*, and *you don't have to*. All these n-grams are highly pervasive in speaking and appear in a wide range of speech events, as the high numbers in the two "Range"

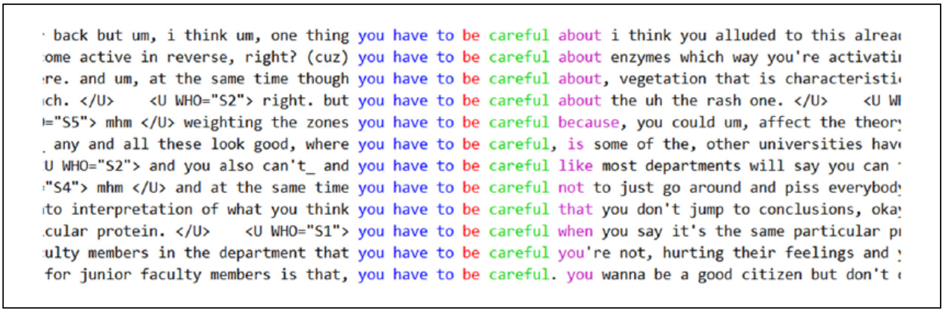
Table 2. The 20 most frequent 3-grams in MICASE and BNC_spoken.

Rank	MICASE			BNC_spoken		
	3-gram	Frequency	Range (number out of 152 texts)	3-gram	Frequency	Range (number out of 915 texts)
1	a lot of	1199	137	a lot of	4513	667
2	I don't know	1107	131	I don't know	3932	584
3	one of the	754	135	I don't think	2943	553
4	you have to	610	125	one of the	2549	608
5	this is the	566	127	do you want	2013	329
6	a little bit	542	130	what do you	1910	430
7	this is a	498	132	and I think	1813	466
8	in terms of	480	104	you want to	1745	460
9	I don't think	473	102	be able to	1656	501
10	I think that	410	86	the end of	1618	512
11	be able to	391	116	you have to	1609	442
12	and this is	390	114	it was a	1599	457
13	you have a	377	111	do you think	1544	394
14	you know what	377	99	I think it's	1538	462
15	I mean I	363	81	a bit of	1537	452
16	part of the	338	103	going to be	1535	435
17	some of the	331	105	I think it	1524	484
18	you can see	326	86	I mean I	1491	381
19	you need to	326	89	I think that	1377	393
20	you know the	321	96	there was a	1360	437

columns in Table 2 indicate. These items are likely to be frequent in spoken English because they serve important roles in the discourse. They function variably as quantifiers (e.g., *a lot of (the), one of the, a bit of*), discourse markers (e.g., *you know, I mean*), and discourse structuring devices (e.g., *in terms of, at the same time*). Other functions expressed by some of the most frequent spoken n-grams identified here are evaluation or stance (e.g., *I don't think, I think that, I think it, you know what I, you can see, I don't know if*), and making (strong) suggestions (e.g., *you need to, you have to, if you look at, have a look at*). All of these appear to be important functions in spoken discourse.

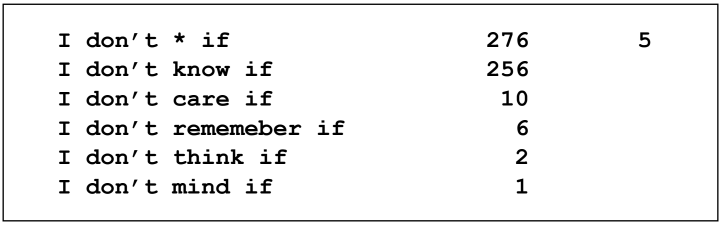
An examination of the concordances of these and other MICASE and BNC_spoken n-grams indicates that the spoken language patterns are actually even more pervasive and more extended than the n-gram lists indicate. As the MICASE concordance sample in Figure 1 shows, the 3-gram *you have to* helps form the larger pattern *you have to be careful (about)*. The same 3-gram is also part of the repeatedly used 6-grams *all you have to do is* and *you have to be able to*. These observations on the extensive patterning of spoken language lend empirical support to Leech's (2000, p. 697) claim that "conversational grammar has a restricted and repetitive lexicogrammatical repertoire".

An additional analytic step carried out to highlight aspects of spoken lexicogrammar involved the extraction of phraseological items that are not entirely fixed (like n-grams)



```
· back but um, i think um, one thing you have to be careful about i think you alluded to this alrea
ome active in reverse, right? (cuz) you have to be careful about enzymes which way you're activati
re. and um, at the same time though you have to be careful about, vegetation that is characteristi
ich. </U>    <U WHO="S2"> right. but you have to be careful about the uh the rash one. </U>    <U WH
="S5"> mhm </U> weighting the zones you have to be careful because, you could um, affect the theor
. any and all these look good, where you have to be careful, is some of the, other universities hav
U WHO="S2"> and you also can't_ and you have to be careful like most departments will say you can
"S4"> mhm </U> and at the same time you have to be careful not to just go around and piss everybod
to interpretation of what you think you have to be careful that you don't jump to conclusions, oka
cular protein. </U>    <U WHO="S1"> you have to be careful when you say it's the same particular p
ulty members in the department that you have to be careful you're not, hurting their feelings and
for junior faculty members is that, you have to be careful. you wanna be a good citizen but don't
```

Figure 1. Part of a right-sorted concordance of the 3-gram *you have to* in MICASE.

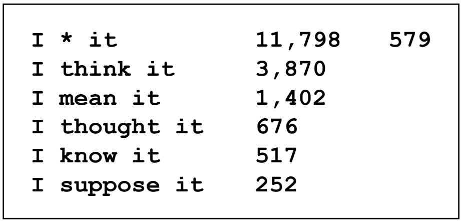


```
I don't * if              276      5
I don't know if           256
I don't care if            10
I don't rememeber if        6
I don't think if            2
I don't mind if             1
```

Figure 2. The phrase-frame *I don't * if* with its five variants in MICASE.

but allow for internal variation. The software *kfNgram* was used to derive so-called phrase-frames (or p-frames), that is, sets of n-grams which are identical except for one word in the same position, from corpora. To give an example, in *kfNgram* the 4-grams *on the one hand* and *on the other hand* would be summarized under the phrase-frame *on the * hand*, with the variants *one* and *other* occupying the variable * slot. P-frames provide insights into pattern variability and indicate to what degree language items are fixed. Variants of a phrase-frame are usually members of the same word class and often form semantically coherent sets. The variants that fill the * slot in *I don't * if* in Figure 2, for instance, are all verbs of cognition, with the most frequent one by far being *know*. Similarly, the most frequent variants of the p-frame *I * if*, a highly frequent spoken frame, as evidenced by our BNC_spoken data (see Figure 3), are cognition verbs, including *think, know*, and *mean*. Another example of a p-frame that is very common in speech (occurring 22,109 times in BNC_spoken) is *a * of*. Its most frequent variants are the nouns *lot, bit, couple*, and *number*. This p-frame allows us to summarize several of the top 3-grams identified in the previous analytic steps under one item. The same applies to the p-frame *you * to* (11,352 instances in BNC_spoken), which covers the high-frequency n-gram variants *you want to, you have to*, and *you need to* (among others), all of which serve a similar pragmatic purpose.

Results from this analysis demonstrate that spoken language patterns are not only extensive and pervasive but also variable – perhaps more variable than a simple n-gram extraction suggests. An additional step in exploring the lexicogrammar of speech (not included here) could be to examine even more variable types of structures, so-called



```
I * it          11,798   579
I think it      3,870
I mean it       1,402
I thought it    676
I know it       517
I suppose it    252
```

Figure 3. The phrase-frame *I * it* with its five most frequent (out of 579) variants in BNC_spoken.

"concgrams" (Cheng, Greaves, & Warren, 2006; Cheng, 2008), which constitute flexible word association patterns that allow for positional and constituency variation. For instance, a concgram search for the associated items "at", "end", and "of" would retrieve all sentences that contain these items in any order and with any number of intervening words. Such a concgram analysis can provide important insights into language patterning but is computationally demanding and involves a substantial amount of manual post-processing of the initial search results. Other analytic techniques in corpus linguistics that enable insights into phraseological aspects of language but which require at least a basic level of programming knowledge include CollGram Analysis (Bestgen & Granger, 2014; Granger & Bestgen, 2014) and Collostructional Analysis (e.g., Stefanowitsch & Gries, 2003; Wulff & Römer, 2009).

Aspects of speech and speaking assessment

The MICASE and BNC_spoken explorations discussed in the previous sections have highlighted words and phrasal sequences that appear frequently in authentic speech in two domains of language use: academic and general non-academic domains. The generated frequency word and (more so) keyword lists point to the centrality of personal pronouns, hesitation markers, short forms, backchanneling and cohesive devices, and discourse markers in spoken language.[2] Insights into some of the core communicative functions of spoken discourse were gained through an examination of lists of phraseological items (n-grams and p-frames) which were found to express quantification, stance or evaluation, used to make suggestions, and help structure and organize the discourse. Phraseological items were shown to be extremely pervasive in general and academic spoken English, confirming observations made by Biber et al. (1999), Carter and McCarthy (2006), and Erman and Warren (2000). According to Erman and Warren's analysis of data from the London-Lund Corpus, prefabricated language patterns account for about 59% of spoken English (2000, p. 37).

The high frequencies of n-grams and p-frames in our MICASE and BNC_spoken data, as well as their extended patterns when looked at in context, point to a close connection between lexis and grammar in spoken language. Words clearly have preferred patterns of occurrence, and structures do not select vocabulary items randomly but in systematic ways. The resulting fixed or semi-fixed lexicogrammatical patterns represent meaningful units that we use to communicate. Echoing Sinclair (2008) quoted above, the analysis of n-grams and phrase-frames demonstrate that the phrase, rather than the word, is the central unit of meaning. Many of the n-grams listed in Table 2 serve to express important discourse functions. Straddling the boundary of lexis and grammar (or vocabulary and syntax), phrases or patterns are the building blocks of spoken discourse and form the core of the construct of speaking. I will now discuss how far these central aspects of speech, including the extensiveness of language patterning and the interconnectedness of lexis and grammar, are captured in the rating scales of the speaking components of the following internationally recognized tests: the TOEFL iBT (internet-based test), CaMLA's ECCE and ECPE, the IELTS, and the Cambridge English: Advanced exam.

TOEFL iBT

The first assessment rating scales under analysis are the integrated and independent speaking rubrics of the TOEFL iBT offered to learners who wish to study at US universities. It assesses test takers' English proficiency for academic and general purposes. The TOEFL iBT scores are predominantly used to gauge "the ability of international students to use English in an academic environment" (Educational Testing Service, 2010, p. 2). The speaking portion of the TOEFL iBT consists of two independent tasks and four tasks that test integrated skills (speaking combined with reading and/or listening).[3]

Starting with the scoring standards for the integrated tasks, notes in the scoring category "Language Use" clearly indicate a strict separation of vocabulary and syntax. Test takers who receive a score of 4 (the highest possible score) demonstrate "good control of basic and complex *grammatical structures*" and their response "[c]ontains generally effective *word choice*" (my emphasis in all quotes in this section). At the score 3 level, a test taker's response may show some "inaccurate use of *vocabulary* or *grammatical structures*". The same separation exists in the descriptions for score levels 1 and 2. The scoring standards for the two independent tasks are very similar in that they treat "*grammar* and *vocabulary*" or "*vocabulary* and *grammatical structures*" separately across score levels. It is interesting to note that the use of formulaic language – a core feature of authentic speech – is highlighted as a feature of the lowest score level where the rubric states that "[s]ome low-level responses may rely heavily on practiced or formulaic expressions". The idea behind this may be to penalize test takers who memorize chunks for the purpose of testing while not necessarily knowing how to use them, but it may still send the wrong signal in implying that formulaic expressions are a feature of low-proficiency English. No mention was found in these two rating scales of any of the aspects of speech determined through our spoken corpus analysis.

CaMLA's ECCE and ECPE

Vocabulary and syntax are also treated as clearly distinct components of language in the rating scales of CaMLA's Examination for the Certificate of Competency in English (ECCE) and their Examination for the Certificate of Proficiency in English (ECPE).[4] The ECCE is a test of "high-intermediate competence in English for personal, public, educational, and occupational purposes", whereas the ECPE provides evidence of a learner's advanced English language proficiency, assessing "linguistic, discoursal, sociolinguistic, and pragmatic elements of the English language" (Cambridge Michigan Language

Assessments, 2014). The speaking portions of both tests consist of interviews in which the test takers participate in a structured (or semi-structured) task with one or two examiners. The ECCE speaking rating scale contains a category called “Language Control & Resources”, which distinguishes between “Grammar” and “Vocabulary” phenomena (listed in separate columns) across scoring levels. However, at each level in the rating scale there is a comment on the presence or absence of “cohesive devices” in a box that spans the “Grammar” and “Vocabulary” columns, signaling an awareness of the importance of such devices in spoken interactions.

Similar to the ECCE, the “Linguistic Resource” category in the ECPE speaking rating scale lists “Grammar” and “Vocabulary” as separate components to assess. The ECPE rating scale does, however, acknowledge the existence of language patterns in speech by including comments on “[c]ollocations, colloquial language, [and] idiomatic expressions”. Such phraseological phenomena are mentioned as characteristic of the performance of higher level test takers. Coherence is included as an aspect of proficient test taker speech in the “Discourse and Interaction” category of the ECPE.

IELTS and Cambridge English: Advanced (CAE)

The final two rating scales included in our survey are those of the International English Language Testing System (IELTS) and the Cambridge English: Advanced exam (CAE), both managed (or co-managed in the case of the IELTS) by Cambridge English Language Assessment. The target audience of the IELTS consists of “people who want to study or work where English is used as the language of communication” (IELTS, 2014). The IELTS speaking component consists of a structured interview between a test taker and an examiner. Similar to the tests previously discussed, the rating scale of the IELTS treats lexis and grammar separately in assessing speaking by defining the distinct band descriptors “Lexical Resource” and “Grammatical Range and Accuracy”.[5] Based on this rating scale, IELTS raters are asked to score a candidate’s appropriate vocabulary use independent of his or her use of grammatical structures. Despite this clear division, the IELTS rating scale still captures some aspects of spoken language patterning. Included under “Lexical Resource” are statements related to whether or not a test taker “uses *idiomatic language* naturally and accurately”, and whether he or she “shows some awareness of style and *collocation*”. A separate “Fluency and Coherence” band descriptor includes comments on a test taker’s appropriate use of “cohesive features” and “discourse markers”, both of which our corpus analysis has identified as being central in authentic speech.

The CAE is “an in-depth assessment of English for people who want to use English in demanding work and study situations” (Cambridge English, 2014). Its spoken component is a structured interview between two candidates and two examiners. Like IELTS, the CAE’s speaking rating scale has a “Lexical Resource” category that is separate from “Grammatical Resource”.[6] The rating scale indicates that appropriate vocabulary use is assessed independently of the use of various grammatical forms. Core aspects of speech, such as the use of cohesive devices and discourse markers, are, however, mentioned under the “Discourse Management” category of the CAE rating scale. Examples of discourse markers and cohesive devices are given in the CAE glossary that accompanies the rating scale. They include several frequent 2-, 3-, and 4-grams (e.g., *you know, as a* *result, on the other hand*), some of which have been discussed in our corpus analysis above. This indicates an awareness of the phrase as the central meaningful unit in language. Also covered in the CAE glossary are the concepts of collocation and fixed phrases (included in the glossary entries for "Range" and "Appropriacy of Vocabulary"). The collocation "big snow" is given as an example of inappropriate vocabulary use together with the more appropriate phrase "heavy snow". The entry for "Range" notes that test takers at higher levels "will make increasing use of a greater variety of words, fixed phrases, collocations and grammatical forms."

From a corpus linguist's perspective, this last rating scale, together with its detailed glossary, appears to better mirror the reality and better capture phraseological features of speaking than any of the other rating scales I examined. The CAE rating scale is more in line with the corpus evidence than the rating scales of the TOEFL iBT, ECCE, ECPE, and IELTS. It does, however, still list grammar and vocabulary as separate components of (spoken) language and does not fully reflect the inseparability of the two attested in this study.

Conclusion and outlook

The aims of this article have been to examine the construct of speaking from a corpus perspective and to link observations on spoken English lexicogrammar to the current practice of rating speaking. To meet these aims, frequent words, keywords, and phraseological items were extracted from two corpora of spoken English (MICASE and BNC_spoken) and analyzed to identify core aspects of authentic speech. The results from the corpus analyses then formed the basis of an examination of rating scales of five different internationally recognized speaking tests. The MICASE and BNC_spoken analyses highlighted the frequent use of personal pronouns, hesitation markers, short forms, discourse markers, and backchanneling and cohesive devices in speech. They also showed that spoken language is dominated by phraseological items which perform central discourse functions and integrate lexis and grammar in a way that makes them inseparable, lending further support to existing statements on lexis–grammar interrelatedness in the corpus linguistics literature (Biber, 2009; Ellis, Römer, & O'Donnell, 2016; Hoey, 2005; Hunston, 2002; Hunston & Francis, 2000; Römer, 2005, 2009, 2010; Sinclair, 1991, 2004, 2008; Stubbs, 2001) and stressing the importance of phraseology as a core, rather than a peripheral aspect of language (cf., Ellis, 2008).

The review of spoken rating scales indicated that the main aspects of spoken English lexicogrammar, as identified in the corpus analysis, are not yet fully reflected in speaking assessment practice. In line with Bachman and Palmer's model of language ability, all examined rating scales include separate scoring categories for lexis or vocabulary on the one hand and grammar or structure on the other, instead of integrating them. Attempts are made to reflect aspects of authentic speech, for example by acknowledging the importance of collocations, cohesive devices or discourse markers, but the major split between "vocabulary" and "grammar" in assessment criteria seems to remain. I would argue that non-holistic rating scales ought to acknowledge the intersection of grammar and vocabulary more explicitly and avoid giving separate scores for two skills that are so closely interrelated. As will be clear from my comments in the present article, I share the concern voiced by Alderson and Kremmel (quoted in the introduction to this article) about the questionable usefulness of treating grammar and vocabulary separately in testing learners' reading ability. I have discussed the same concern here with reference to assessing speaking. I believe that, if the goal is to bring speaking assessment more in line with the reality of authentic speech, a more phraseological approach may be needed to reflect the pervasive lexicogrammatical patterning of speech in rating scales of speaking tests.

The data presented in this paper have additional implications for testing practice, as research suggests that raters frequently have difficulty in distinguishing vocabulary and grammar in the assessment of test taker performance. A study by Ruegg, Fritz, and Holland (2011) shows that raters of the Kanda English Proficiency Test are finding it difficult to keep lexis and grammar apart and often give the same score for both areas (see also Ruegg, 2015). Their findings suggest that an integrated approach that allows raters to give one score for lexicogrammar instead of two separate ones may facilitate the practice of scoring tests. Although their focus is on assessing writing, Ruegg et al.'s (2011) arguments may be transferable to the context of speaking assessment. A possible adjustment to rating scales would hence be to merge categories such as "Lexical Resource" and "Grammatical Resource" to create a "Lexicogrammatical Resource" category. In the rating scale of CaMLA's recently revised International Teaching Assistant Speaking Assessment (ITASA), lexicogrammar is covered under the category "Transactional Competence".[7] This category still refers to test takers' use of "vocabulary and grammar", but captures the importance of lexis-grammar inseparability by including the use of "nativelike formulaic sequences" (phrases such as "be unaware of" or "on the other hand") as an aspect of the highest level of proficiency.

Although I recommend the revision of existing rating scales in the light of corpus evidence on the patterned nature of speaking, I also see a need for more research that highlights aspects of speech, works towards operationalizing "phraseology" or "lexicogrammar" (cf. O'Donnell, Römer, & Ellis, 2013 on measuring formulaicity), and makes concrete suggestions on how to implement a data-driven, phraseological approach in rating scale development. As a corpus linguist, I do not have the relevant expertise to make such concrete suggestions but would need to collaborate with assessment experts. Future research also needs to discuss the effectiveness of revised assessment scales and demonstrate that revising rating scales in the way suggested here will lead to more valid ratings. I believe that collaborative work between corpus linguists and testing specialists will enable us to address these issues, gain a better understanding of real-world spoken lexicogrammar, and allow us to increase the level of authenticity in speaking assessment.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. The software settings in *AntConc* were changed to include punctuation in the token definition so as to keep word-internal apostrophes and not to split contracted forms such as "don't" into "don" and "t". Commas and periods (representing mid-utterance and utterance-final pauses respectively) hence appear as part of items in the frequency word and keyword lists.
2. Frequency word and keyword lists based on discipline- or domain-specific language (e.g., lectures in Engineering, or Business English presentations) would, of course, also be useful tools if one wanted to identify domain-specific vocabulary.
3. TOEFL iBT rating scales were retrieved from www.ets.org/Media/Tests/TOEFL/pdf/Speaking_Rubrics.pdf.
4. ECCE and ECPE rating scales were retrieved from www.cambridgemichigan.org/sites/default/files/resources/rating-scales/ECCE-Rating-Scale-Speaking-20140220.pdf and http://www.cambridgemichigan.org/sites/default/files/resources/rating-scales/ECPE-RatingScale-Speaking.pdf.
5. The rating scale for the speaking component of the IELTS was retrieved from www.ielts.org/pdf/Speaking%20Band%20descriptors.pdf.
6. The CAE speaking scale was retrieved from www.teachers.cambridgeesol.org/ts/digitalAssets/117408_CambridgeEnglish_Advanced__CAE__Handbook.pdf.
7. The ITASA rating scale was retrieved from www.cambridgemichigan.org/sites/default/files/resources/itasa/ITASA-RatingScale.pdf.

References

Alderson, J. C., & Kremmel, B. (2013). Re-examining the content validation of a grammar test: The (im)possibility of distinguishing vocabulary and structural knowledge. *Language Testing*, *30*(4), 535–556.

Anthony, L. (2014). *AntConc (Version 3.4.3)* [Computer Software]. Tokyo, Japan: Waseda University. Retrieved from www.laurenceanthony.net/.

Bachman, L., & Palmer, A. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bachman, L., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, *26*, 28–41.

Biber, D. (1988). *Variation in speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, *14*(3), 275–311.

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

Bloomfield, L. (1933). *Language*. New York: Holt, Rinehart and Winston.

Burnard, L. (2007). *Reference guide for the British National Corpus* (XML edition). Retrieved from www.natcorp.ox.ac.uk/docs/URG/.

Cambridge English (2014). *What's in the exam?* Retrieved from www.cambridgeenglish.org/exams-and-qualifications/advanced/whats-in-the-exam/.

Cambridge Michigan Language Assessments (2014). *Exams*. Retrieved from www.cambridgemichigan.org/exams.

Carter, R., & McCarthy, M. (1995). Grammar and the spoken language. *Applied Linguistics*, *16*(2), 141–158.

Carter, R., & McCarthy, M. (2006). *Cambridge grammar of English: A comprehensive guide*. Cambridge: Cambridge University Press.

Cheng, W. (2008). Concgramming: A corpus-driven approach to learning the phraseology of discipline-specific texts. *CORELL: Computer Resources for Language Learning*, *1*, 22–35.

Cheng, W., Greaves, C., & Warren, M. (2006). From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics*, *11*(4), 411–433.

Council of Europe. (2001). *Common European Framework of Reference for Languages*. Cambridge: Cambridge University Press. Retrieved from www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf.

Educational Testing Service (2010). TOEFL iBT Test Framework and Test Development. TOEFL iBT Research Insight, *1*. Retrieved from www.ets.org/s/toefl/pdf/toefl_ibt_research_insight.pdf.

Ellis, N. C. (2008). Phraseology: The periphery and the heart of language. In F. Meunier & S. Granger (Eds.), *Phraseology in language learning and teaching* (pp. 1–13). Amsterdam: John Benjamins.

Ellis, N. C., Römer, U., & O'Donnell, M. B. (2016). *Usage-based approaches to language acquisition and processing: Cognitive and corpus investigations of construction grammar*. Malden, MA: Wiley.

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text*, *20*(1), 29–62.

Fletcher, W. H. (2007). *KfNgram*. Annapolis, MD: USNA. Retrieved from www.kwicfinder.com/kfNgram/kfNgramHelp.html.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing*, *13*(2), 208–238.

Fulcher, G. (2003). *Testing second language speaking*. London: Longman.

Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics*, *52*(3), 229–252.

Hoey, M. (2005). *Lexical priming. A new theory of words and language*. London: Routledge.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Hunston, S., & Francis, G. (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.

Hyland, K. (1998). *Hedging in scientific research articles*. Amsterdam: John Benjamins.

IELTS (2014). *What is IELTS?* Retrieved from www.ielts.org/test_takers_information/what_is_ielts.aspx.

Lado, R. (1961). *The construction and use of foreign language tests*. New York: McGraw-Hill.

Leech, G. (2000). Grammars of spoken English: New outcomes of corpus-oriented research. *Language Learning*, *50*(4), 675–724.

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

O'Donnell, M. B., Römer, U., & Ellis, N. C. (2013). The development of formulaic sequences in first and second language writing: Investigating effects of frequency, association, and native norm. *International Journal of Corpus Linguistics*, *18*(1), 83–108.

Römer, U. (2005). *Progressives, patterns, pedagogy: A corpus-driven approach to English progressive forms, functions, contexts and didactics*. Amsterdam: John Benjamins.

Römer, U. (2009). The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics*, *7*, 140–162.

Römer, U. (2010). Establishing the phraseological profile of a text type: The construction of meaning in academic book reviews. *English Text Construction*, *3*(1), 95–119.

Ruegg, R. (2015). An experiment in the ability of raters to evaluate lexis in writing. *Language in Focus Journal*, *1*(1), 38–50.

Ruegg, R., Fritz, E., & Holland, J. (2011). Rater sensitivity to qualities of lexis in writing. *TESOL Quarterly*, *45*(1), 63–80.

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

Shiotsu, T., & Weir, C. J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing*, *24*(1), 99–128.

Sinclair, J. M. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, J. M. (2004). *Trust the text: Language, corpus and discourse*. London: Routledge.

Sinclair, J. M. (2008). The phrase, the whole phrase, and nothing but the phrase. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 407–410). Amsterdam: John Benjamins.

Simpson, R. C., Briggs, S. L., Ovens, L., & Swales, J. M. (2002). *The Michigan Corpus of Academic Spoken English*. Ann Arbor, MI: The Regents of the University of Michigan.

Stefanowitsch, A., & Gries, S. Th. (2003). Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics*, *8*(2), 209–43.

Stubbs, M. (2001). *Words and phrases. Corpus studies of lexical semantics*. Oxford: Blackwell.

Taylor, L. (Ed.) (2011). *Examining speaking: Research and practice in examining second language speaking*. Cambridge: Cambridge University Press.

Vidaković, I., & Galaczi, E. D. (2013). The measurement of speaking ability 1913–2012. In Weir, C. J. Weir, I. Vidaković, & E. D. Galaczi *Measured constructs: A history of Cambridge English language examinations 1913–2012* (pp. 257–346). Cambridge: Cambridge University Press.

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