Focus on ELT Journal (FELT)
Vol 1, Issue 1, 2019 ISSN: 2687-5381

The linguistic dimension of L2 interviews: A multidimensional analysis of native speaker language

[a]Pascual Pérez-Paredes and [b]María Sánchez-Tornel
a Lecturer, University of Cambridge, United Kingdom, pfp23@cam.ac.uk
b EOI San Javier, Spain, mstornel@um.es

To cite this article: Pérez-Paredes, P., & Sánchez-Tornel, M. (2019). The linguistic dimension of L2 interviews: A multidimensional analysis of native speaker language. *Focus on ELT Journal (FELT), 1*(1), 4-26. https://doi.org/10.14744/felt.2019.00008

ABSTRACT

This research profiles L2 interviews from a variationist perspective by using native speaker data in order to gain insight into the characteristics of three different speaking tasks in the framework of the *LINDSEI* learner language corpus tradition: Personal Narrative Component, an Interaction Component and a Picture Description. This way, we set out to research one area of the assessment of proficiency that is usually neglected: that of the linguistic nature of the tasks used to assess general "proficiency" in a given language. Our corpus was part-of-speech (POS) tagged and analysed using Multidimensional Analysis (MDA). We found that the different speaking tasks determine the range of linguistic features that are more likely to be generated by the communicative potential of the task itself. This profiling is of interest in areas such as language assessment, where the interview is widely used to evaluate the speakers' communicative competence, but also in the field of learner language research.

Keywords:
oral proficiency interview
native speaker language
corpus linguistics
Multidimensional Analysis (MDA)
dimensions of language use

Introduction

Interviews have been used extensively as an elicitation technique either for language research (Gilquin & Gries, 2009) or for communicative competence appraisal. Apart from the cue-based interviews used to evaluate the depth of vocabulary knowledge (Kunnan, 1998), interviews in the context of English as a Foreign Language (EFL) are regularly conducted to assess the communicative competence of language learners. International institutions like the *American Council on the Teaching of Foreign Languages* (ACTFL), *Cambridge English for Speakers of Other Languages* (ESOL) *Examinations*, or Trinity College, among others, use the oral proficiency interview (OPI) to test the oral competence of candidates worldwide. In the US, agencies such as the CIA, the FBI, and the DLI have been using L2 interviews to assess the foreign language speaking capabilities of their employees since the 1950s (Johnson 2001, p. 7).

Given the tradition of assessing learner language by means of interviews, it is hardly surprising that the interview has been the most widely used elicitation technique in the collection of spoken learner data (Tono, 2003). In the field of learner language research, the publication of the first spoken learner corpus, the Louvain International Database of

✉Contact: pfp23@cam.ac.uk

Spoken English Interlanguage (LINDSEI) (Gilquin, De Cock and Granger, 2010), which was compiled by means of oral interviews, was a major breakthrough in the analysis of spoken learner language. The new Trinity Lancaster Corpus (TLC) (Gablasova, Brezina and McEnery, 2019) will contribute to our understanding of how L2 English is used in oral proficiency interviews across a variety of tasks and, interestingly, performance levels.

Despite its importance in learner language research and learner language assessment, the L2 interview as a linguistic register remains under-researched. Iwashita, Brown, McNamara and O'Hagan (2008) have pointed out how different authors have tried to gain further insight into the features of the language produced by test-takers (Shohamy, 1994), the speech event(s) in L2 interviews (Van Lier, 1989), or the relations between candidates' performance and the scores awarded (McNamara et al., 2002). Given the widespread use of interviews and the lack of research in this register from a native speaker perspective, we set out to gain insight into the nature of L2 interviews through Multidimensional Analysis (MDA). Specifically, we want to find out whether the three speaking tasks that were used to gather our corpus can be profiled distinctively**.** If so, what other registers do these sub-registers resemble? In this research, we aim at profiling the L2 interview from a variationist perspective, using English native speaker data in order to shed light into the characteristics of this particular register as manifested across different speaking tasks. We argue that tasks do not just simply prompt different language use, but they actually afford the usage of a set of specific linguistics features.

L2 interviews: speaking tasks, language assessment and corpora

Corpora in the assessment and operationalization of proficiency

Corpus-based approaches are widely considered as central to diverse areas of language study including, among others, Language Testing and Assessment (LTA). This field has, for many years now, benefited from the use of real language data in various respects. Alderson (1996) presented one of the first accounts of the potential uses of corpora in language assessment. Among these we find test construction, compilation and selection, test presentation, response capture, test scoring and calculation and delivery of results. Given the limited use of computers (let alone language corpora) in language testing at the time, the author referred to his account as mere speculation, but he anticipated that "since corpora exist, they will eventually be used, for better or worse […] it makes sense to think about how to best use them in order to control their development rather than to suffer it" (Alderson, 1996, p. 249). Time has proven that Alderson was not far wrong, as the following paragraphs will illustrate.

Since the creation of the *Cambridge Language Corpus* (Cambridge ESOL Examinations) as a repository of rubrics and exam answers transcripts, the use of native speaker and learner corpora in LTA has unfolded in different directions. The application of corpus methods to analyze native speaker or learner data is indeed valuable in LTA, whether it be aimed at profiling and characterizing proficiency, at assessing it or at informing and validating test design. In the language testing tradition, native speaker (NS) and non-native speaker (NNS) corpora have been used to revise tests, devise new test formats as well as teaching and testing materials, create and/or revise wordlists, shed light on the characteristics of academic speech and investigate differences by speaker group or discipline (Taylor and Barker, 2008; Barker, 2010). Native speaker data help to make sound decisions on structures, phrases or vocabulary which are to be included or avoided in tests, thus leaving test writers' intuitions and experiences out of the picture (Barker, 2010). Furthermore, they serve as a source of real-life texts that can be adapted or used without further editing and also as a reference resource in the stage of marking or grading. As for learner corpora, they have been used, among other aspects, to identify what learners can do and the errors that are common at a given proficiency level, to confirm test writers' intuitions about the features that are typical of certain levels, to revise rating scales, to explore automatic rating, or to analyze the relationship between demographic variables, test mode and learning environment on learner output (Barker, 2010; Taylor and Barker, 2008).

Much as the use of corpora has resulted in the advancement and improvement of LTA, it is no less true that the definition of proficiency and the delimitation of the boundaries of different proficiency bands still seem to be rather challenging for test designers and Second Language Acquisition (SLA) researchers alike. Carlsen (2012, p. 162) has pointed out that "levels of proficiency are not always carefully defined, and the claims about proficiency levels are seldom supported by empirical evidence" and Barker (2010) highlights that "establishing the nature of language proficiency at different levels is vital for language testers seeking to design tests that either aim to assess candidates at a particular proficiency level or report results across part of or the whole proficiency scale." Moreover, the correct placement of learner corpus texts in their corresponding proficiency bands has further implications, given that the linguistic features expected in those bands can only be isolated reliably "if a learner's level is correctly identified and recorded in a corpus" (Barker 2010, p. 637). The importance of ensuring validity and reliability with respect to the assignment of learner corpus texts to different levels of proficiency stands out, therefore, as a shared concern in the field, since erroneous decisions may lead SLA experts to make spurious assumptions regarding language learning. Díez-Bedmar (2018, p. 208) has "highlighted the main challenges that linguistic competence descriptors pose to CEFR and ELP users […] with a particular focus on the grammatical accuracy descriptors and strategy descriptors for monitoring and repair at B1 level". From this it follows that a sound approach to LTA depends greatly on a series of factors that are closely interwoven, ranging from the precise characterization of proficiency and proficiency levels to the right design of tests, all of them informed and supported by NS and NNS corpora in different ways.

Bearing these concerns in mind, one might go one step further and question the validity of certain tasks that are commonly found in language tests and, in particular, in the speaking section of language tests. This is an area that has not attracted much attention to date as tasks oriented towards the assessment of specific linguistic features may or may not actually bring to the surface the use of such linguistic features, even when NS perform these oral texts.

The study of the potential of tasks to elicit the use of particular phrases, structures or vocabulary that may, presumably, be produced while solving those tasks seems a promising area within LTA. It has not yet been established whether specific tasks are as adequate as test designers expect them to be and it is precisely here where the analysis of native speaker language by means of MDA advocated in this study may play a central role. MDA of learner language has been underused as a tool for language research and pedagogy. One of the few studies where MDA was used to explore learner language is Connor-Linton and Shohamy (2001) and one of the few pedagogic applications of MDA is Aguado et al. (2012). Considering that corpus techniques have proved useful in the analysis and characterization of learner output and in the exploration of native speaker language oriented towards test design and validation, it remains to be seen how LTA and learner corpus research (LCR) can benefit from the study of L2 interviews from a variationist perspective by using MDA.

We adopt, therefore, a critical perspective on task and test design and propose the use of MDA to examine the potential of the L2 interview to elicit an adequate and sufficient number of linguistic features. The underlying principle is that it cannot be assumed that a task is valid or reliable to assess oral proficiency in the light of the presence of particular features without knowing, first, if those features would be employed by a native speaker performing the same task. The application of MDA in LTA is mainly based on the works carried out by Douglas Biber. In the *TOEFL 2000 Spoken and Written Academic Language (T2K-SWAL) Project*, Biber and his colleagues brought together MDA and NS corpora to investigate the linguistic characteristics of institutional registers at university and thus "ensure that the texts used on listening and reading exams accurately represent the linguistic characteristics of spoken and written academic registers" (Biber et al., 2004, p. 2). In a previous investigation Biber and Jamieson (1998, cited in Taylor and Barker, 2008) found that the reading and listening texts did not fully match the registers being tested, which calls for a closer examination of language tasks in the light of MDA.

The L2 interview and language learner assessment

The L2 interview is the "dominant approach to measuring a language learner's oral proficiency" (Connor-Linton and Shohamy, 2001, p. 124), being widely used nowadays (Ricardo-Osorio, 2008) by different and prestigious institutions (Ferrara, 2008). Cambridge ESOL[1] runs different examinations which target a wide spectrum of levels. The *First Certificate of English* (FCE) examiners run an oral test to "assess the candidate's ability to produce spoken English in a variety of tasks". This test involves two candidates and two examiners. The first part of the oral test is an interview where the interlocutor asks each candidate questions which "relate to [his or her] own lives and focus on areas such as work, leisure time, future plans" and social language. The second part of the test is an "individual long turn" where the candidates have to fulfill a one-minute speaking task where two photographs are shown and a printed question has to be answered. This part "tests the candidate's ability to produce an extended piece of discourse which may involve comparing, describing and expressing opinions". The third part of the test, labeled collaborative task, is a "two-way discussion between the candidates, developed around a topic-based visual stimulus" where the candidate's ability to sustain an interaction, exchange ideas, express and justify opinions, agree and/or disagree, make suggestions,

[1] http://www.cambridgeesol.org/assets/pdf/fcecae_review10.pdf

speculate, evaluate and work towards a negotiated outcome is evaluated". Finally, a discussion on one of the topics in the third part is promoted by the interlocutor so as to evaluate the candidate's ability to "engage in a more in-depth discussion, exchange information, express and justify opinions and agree and/or disagree". In total, the test runs for approximately 14 minutes and involves personal information, description of visual prompts and the expression of ideas and opinions over a given topic.

The *Cambridge Advanced English* (CAE) test follows an identical format, while the *Cambridge Certificate of Proficiency* in English (CPE) extends a little longer and may last up to 19 minutes. However, the structure of the interview and its distribution is almost identical: an interview and a collaborative task followed by a discussion between two candidates, one assessor who remains silent, and an interlocutor. The *ACTFL Oral Proficiency Interview* is a standardized procedure for the global assessment of functional speaking ability. It is a face-to-face or telephone interview between a certified ACTFL tester and an examinee that determines how well a person speaks a language by comparing his or her performance in specific communication tasks with the criteria for each of ten proficiency levels described in the *ACTFL Proficiency Guidelines-Speaking*.

The use of interviews is not restricted to the evaluation of General English. Trinity College London runs *Spoken English for Work* (SEW) examinations which "address [a] growing demand [of use of spoken English in real work settings] by offering a face-to-face assessment which measures spoken English in a working context relevant to the chosen profession of the candidate". The four levels range from B1 to C1 and take from 13 to 27 minutes. In all of them, one-to-one, face-to-face assessment is involved, including a telephone task and a topic discussion led by the examiner. Interactive tasks are present in all levels except for B1 and topic presentations are evaluated in the two higher level.

L2 interviews and speaking tasks

L2 researchers have addressed the effect of the speaking task on the linguistic nature of L2 interviews from, at least, two different perspectives. First, we find research which has analyzed the interview as a register. Second, there is research which has limited its scope to discrete linguistic elements. Connor-Lynton and Shohamy (2001) studied the stylistic variation of NNS' spoken discourse across different elicitation tasks and contexts (face-to-face vs taped-mediated). Using MDA, the authors analyzed the data in Shohamy, Donitsa-Schmidt and Waizer (1993), viz. 10 adult female L1 Hebrew EFL learners of varied proficiency levels. These individuals completed three different tasks in parallel forms in order to minimize memorization effects. In the first, they told their interviewer about themselves; in the second, using the role-play technique, they were asked to complain about noise; in the third, they had to request of a professor an extension on a term paper or a second chance on a final exam. These tasks were combined with five elicitation contexts (face-to-face conversation with a tester, with a peer, telephone interaction, videotaped prompt and audio taped prompt). The authors found that the *t*-tests of the dimension scores confirmed that each pair elicited "stylistically and functionally equivalent performance samples" (Connor-Lynton and Shohamy, 2001, p. 133). Similarly, their MDA analysis provided evidence that the stylistic profiles of complaints and requests elicited similar language in terms of communicative functions, which, according to the authors, shows some of the potential uses of MDA in designing L2 interviews which can discriminate a more varied set of speech events.

Johnson (2001) attempted to characterize the L2 interview in terms of speech events through a discourse analysis methodology. The data that the author used were 35 telephone interviews codified according to five major categories, namely, floor turn, repair, topic, question type and discourse unit. Her analysis concludes that the L2 interview resembles more accurately a monologic speech event, rather than conversation. Neary-Sundquist (2009) examined the relationship between the effect of proficiency levels and task types on the use of cohesive devices in English and German second language speech production under test conditions that followed the ACTFL. In the German data, the narrative task showed a higher frequency of use of conjunctions and a decrease in discourse marker use. In the English data, the leaving-a-telephone-message task behaved significantly different from the other tasks as to the frequency of discourse markers. The author concludes that the degree of structure in a task may have an impact on language performance.

Methodology

Corpus used in the analysis

The corpus used in this analysis is the extended LOCNEC (*Louvain Corpus of Native English Conversations* (LOCNEC) (Pérez-Paredes and Bueno, 2019). The LOCNEC (De Cock, 2004) is made up of 90,300 words contributed by 50 native speakers of English, all of them undergraduate and graduate students at Lancaster University. The extended LOCNEC includes 28 extra interviews from the British component of the CAOS-E corpus (Aguado et al., 2012). It is made up of 21,509 words contributed by British undergraduate students at Manchester Metropolitan University.

The extended LOCNEC was compiled following the same format of The *Louvain International Database of Spoken English Interlanguage* (LINDSEI; De Cock, 1998; Gilquin, De Cock and Granger, 2010). First, informants were given three topics for discussion, i.e., an experience that has taught the interviewee an important lesson, a country that has made an impression on the interviewee or a film or play that has attracted their attention. Then, the interviewer engaged the interviewee into an even more involved, interpersonal communication by asking about their studies or future plans. In the last part of the interview, the interviewee was given four pictures that represented a story and was asked to offer an account of what was going on. Table 1 summarizes the main characteristics of the corpus.

Table 1. *Characteristics of the extended LOCNEC (Pérez-Paredes & Bueno, 2019)*

Number of speakers	78
Nationality	British
Interview locations	Lancaster University and Manchester Metropolitan University
Running words	111,809
Speaking tasks/ Components	Personal Narrative Interaction Picture Description

The first part of the interview gave the speaker the opportunity to build a narrative based on his own previous life, travelling or film-viewer experiences. The second is mainly interactional. The interviewer asks the interviewee questions that provide an occasion for the interviewee to talk about themselves and their activities at the moment when the interview took place. These two parts favour involved production. As regards the third part of the interview, the picture description task offers speakers the possibility to elaborate on individual interpretations arising from a situation in which a woman is being portrayed by a painter, and where she seems to be dissatisfied with the painter's first piece of work. This last part of the interview can be regarded as description-oriented production.

Analysis

Our interview corpus was POS tagged and analysed using MDA (Biber, 1988; Conrad, 2001; Biber, 2006). This methodology seeks to interpret linguistic data in the light of language variation across registers or different dimensions of use. Each dimension of use "comprises a distinct set of co-occurring linguistic features, and each has distinct functional underpinnings" (Biber, Reppen and Conrad 2002, p. 459). The five dimensions of use in Biber (1988) are (D1) involved versus information production, (D2) narrative versus non-narrative concerns, (D3) explicit versus situation-dependent reference, (D4) overt expression of persuasion and (D5) abstract versus non-abstract information. Accordingly, five dimension scores were computed for each interview and for each of the parts of the interviews in the corpus. After that, a factor score[2] was calculated. All the frequencies were standardized to a mean of 0.0 and a standard deviation of 1.0 before the computation of the factor. Differences between the three components were tested using the Duncan's Multiple Range Tests, a procedure based on the comparison of the range of a subset of the sample means with a calculated least significant range. The analysis of our data followed the guidelines in Biber, Johansson, Leech, Conrad and Finegan (1999) and, in particular, took into account the tasks performed in discourse by the different linguistic features, the processing constraints which the pedagogic interview register presents, and the conventional association of linguistic features with the peculiarities of the interview situation analyzed.

[2] A factor score is a numerical value that indicates a text relative standing on a latent factor in factor analysis.

Results

Each interview in our research corpus was composed of three different speaking tasks, namely, a Personal Narrative Component, an Interaction Component and a Picture Description. Table 2 shows the scores of the three speaking tasks on the five dimensions of language use in Biber (1988) plus the score of the whole interview, that is, the unabridged, complete interview.

Table 2. *Scores of the speaking tasks on the five dimensions of language use in Biber (1988)*

	D1	D2	D3	D4	D5
Personal Narrative	27	-0.7	-4	-2	-2
Interaction	31	-2	-5	0.13	-2
Picture Description	24.6	-0,1	-5	-4	-0.1
Whole interview	29.50	-1.10	-4.70	-1.02	-1.50

In the following paragraphs we will provide the score of the different speaking tasks on these five dimensions together with the normalized counts of the most relevant linguistic features for each of the dimensions of use.

Dimension 1: Involved versus information production

This dimension marks affective or interactional content, as opposed to information density and exact informational content. Its internal composition makes it possible that much of the variability found in texts can be explained using this factor alone, which turns D1 into a fundamental dimension to discriminate textual variation (Biber 1988: 106). The whole interview scored high on this rank (29.5), above the original interview texts (17.01) in Biber (1988). The Personal Narrative Component score on this dimension (27) is closer to face-to-face conversations in Biber (1988) than the Picture Description Component (24.6), which in turn is closer to spontaneous speech and interviews in Biber (1988). This fact can be explained by the presence of fewer turns in this component, with the interviewer mainly offering backchanneling.

The Interaction Component score (31) places this part of the interview on top of this dimension, lying closer to face-to-face conversations than any other speaking task. Figure 1 shows the scores of all three tasks and the interview mean.

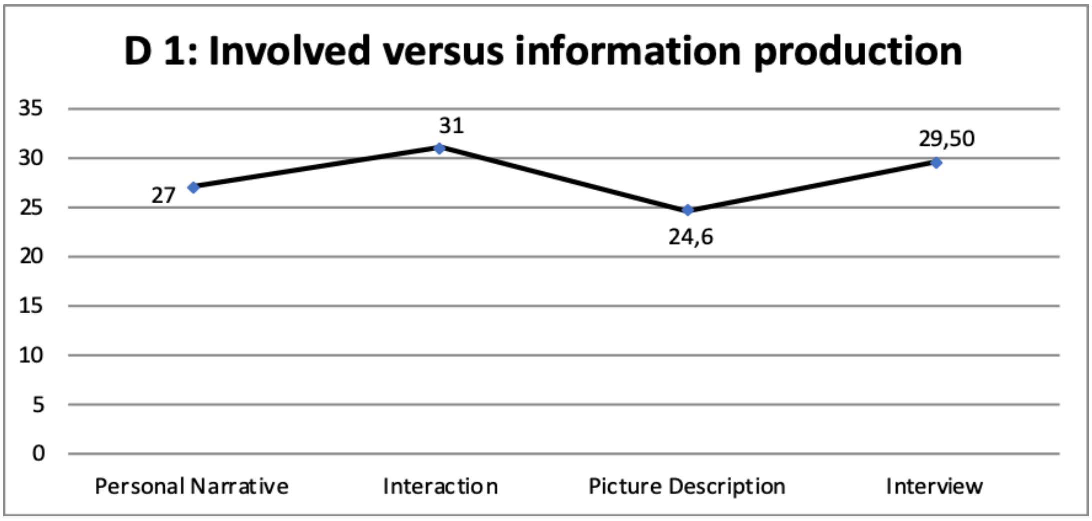

Figure 1. Interview scores on Dimension 1

The score difference between the Interaction Component and the Picture Description (6.4) seems to indicate that the speaking task plays an important role in the way LINDSEI-format interviews can be linguistically profiled. This is confirmed by the Duncan's Multiple Range Test for D1, which shows that the Interaction Component is significantly different from the other two tasks. Table 3 shows the results of the test.

Table 3. *Effect of speaking task on D1 profiling*

Dimension 1: Involved versus information production		
Duncan Grouping	Mean	Speaking task
A	30.830	Interaction
B	27.293	Personal Narrative
B		
B	24.643	Picture Description

Alpha 0.05
Error Degrees of Freedom 226
Error Mean Square 112.6946
Harmonic Mean of Cell Sizes 76.20474

Number of Means	2	3
Critical Range	3.389	3.567

The higher score of face-to-face conversations (35.3) in Biber (1988) seems to point out that our interviews presented fewer opportunities for affectiveness and involvement than conversations, although both registers may share similar real-time production constraints. Spontaneous speech and interviews in Biber (1988) behave in a very similar way on this dimension of use, which confirms that the Involvement Component of our corpus is an efficient register delimiter, at least when compared to the interviews in Biber (1988)[3].

The linguistic features which are representative of the involved dimension include, in decreasing order of significance, private verbs, *that*-deletion, contractions, present tense verbs, 2nd person pronouns, *do* as pro-verb, analytic negation, demonstrative pronouns, general emphatics and 1st person pronouns. Table 4 lists the normalized means of selected features in our corpus.

Table 4. *Summary of SMD estimate across articles with 95% Confidence Interval*

	private verbs	*that*-deletion	contractions	present tense verbs	2nd pers. pronouns	*do* as pro-verb
Personal Narrative	23/1000	7.9/1000	3.0/1000	69.9/1000	23.9/1000	2/1000
Interaction	27.7/1000	10.1/1000	3.3/1000	96/1000	38.6/1000	3.3/1000
Picture Description	16.5/1000	6.7/1000	7.3/1000	118/1000	23.3/1000	1.2/1000
Corpus mean	22.7/1000	8.3 /1000	4.6/1000	97/1000	29/1000	2.3/1000

Other linguistic features are typically representative of information-oriented discourse: nouns, prepositions and attributive adjectives, see Table 5.

Table 5. *Linguistic features which are representative of information-oriented discourse*

	nouns	prepositions	attributive adjectives
Personal Narrative	164.7/1000	74.1/1000	17.2/1000
Interaction	157.6/1000	72.5/1000	15.6/1000
Picture Description	138.2/1000	62.3/1000	10.1/1000
Corpus mean	151.2/1000	68.6 /1000	14/1000

Dimension 2: Narrative versus non-narrative concerns

Dimension 2 distinguishes narrative discourse from other registers where exposition or description are more pivotal. Romantic and mystery fiction texts appear at one end of this continuum, while broadcasts and official documents qualify for a type of text where narration plays a very limited or no role at all (Biber, 1988).

[3] The interviews in Biber (1988) come from the London-Lund Corpus and are classified as part of the public discussion genre.

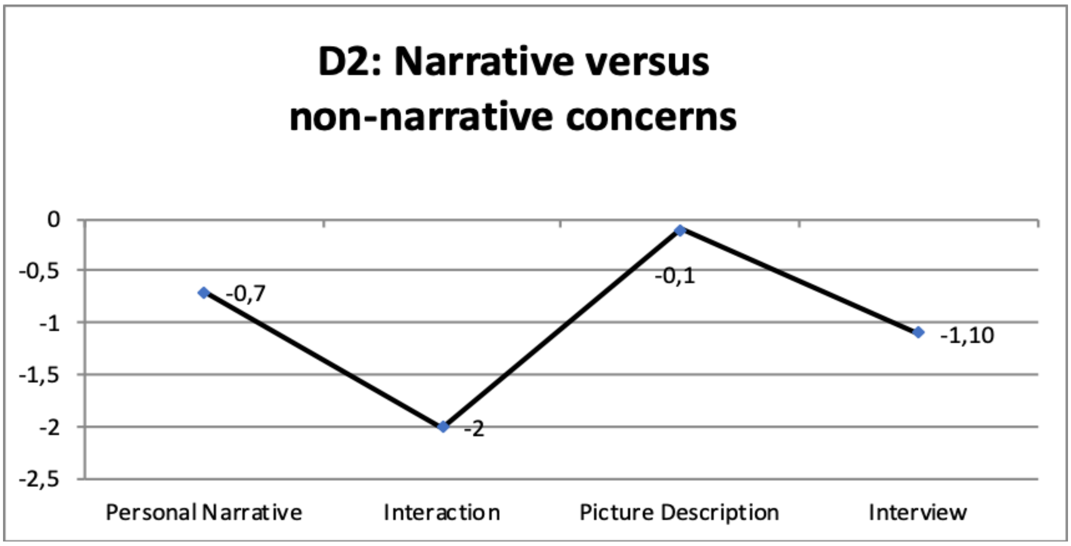

Figure 2. Interview scores on Dimension 2

The mean corpus score (-1.1) is identical to that of the interview texts (-1.1) in Biber (1988). On this dimension, interview texts in Biber (1988) and our corpus data behave exactly in the same way. The interview Personal Narrative Component score on this dimension (-0.7) is close to that of interview texts, and almost identical to that of face-to-face conversations (-0.6) in Biber (1988). The score of the Picture Description Component (-0.1) is slightly farther away from face-to-face conversations, while the Interaction Component score (-2) matches that of telephone conversations in Biber (1988). The score difference between the Interaction Component and the Picture Description Component (1.9) seems to indicate that the speaking task does play an important role in the way interviews can be linguistically profiled. This is confirmed by the Duncan's Multiple Range Test for D2, which shows that all three corpus components are significantly different from each other. Table 6 shows the results of the test.

Table 6. *Effect of speaking task on D2 profiling*

Dimension 2: Narrative versus non-narrative concerns		
Duncan Grouping	Mean	Speaking task
A	-0.1059	Picture Description
B	-0.6947	Personal Narrative
C	-1.6835	Interaction
Alpha 0.05 Error Degrees of Freedom 226 Error Mean Square 3.330233 Harmonic Mean of Cell Sizes 76.20474 Number of Means 2 3 Critical Range .5826 .6132		

The linguistic features which are representative of the narrative dimension include, in decreasing order of significance, past tenses, 3rd person pronouns, perfect aspect tenses and public verbs. Present tenses and attributive adjectives are typical features of non-narrative texts. Table 7 lists the normalized frequencies of these linguistic features in our corpus data.

Table 7. *Effect of speaking task on D2 profiling*

	past tense	3rd pers. pronouns	perfect aspect	public verbs	present tense	attributive adjectives
Personal Narrative	61.9/1000	20.8/1000	8.6/1000	2.3/1000	69.9/1000	17.2/1000
Interaction	32.8/1000	12.3/1000	10.4/1000	2.2/1000	96/1000	15.6/1000
Picture Description	8.8/1000	94.5/1000	2.7/1000	3.4/1000	118/1000	10.1/1000
Corpus mean	31.9/1000	46.2/1000	6.8/1000	2.8/1000	97/1000	14/1000

Dimension 3: Explicit versus situation-dependent reference

This dimension distinguishes between discourse which identifies referents in an explicit way, mainly through relatives, from discourse that relies more heavily on non-specific deictics (Biber 1988: 115). The score of our interview corpus (-4.7) is far away from that of interviews (-0.4) and spontaneous speeches (1.2) in Biber (1988). On this dimension, our corpus behaves in a similar way to telephone (-5.2) and face-to-face conversations (-3.9). Figure 3 shows the scores of the components of our corpus on dimension 3.

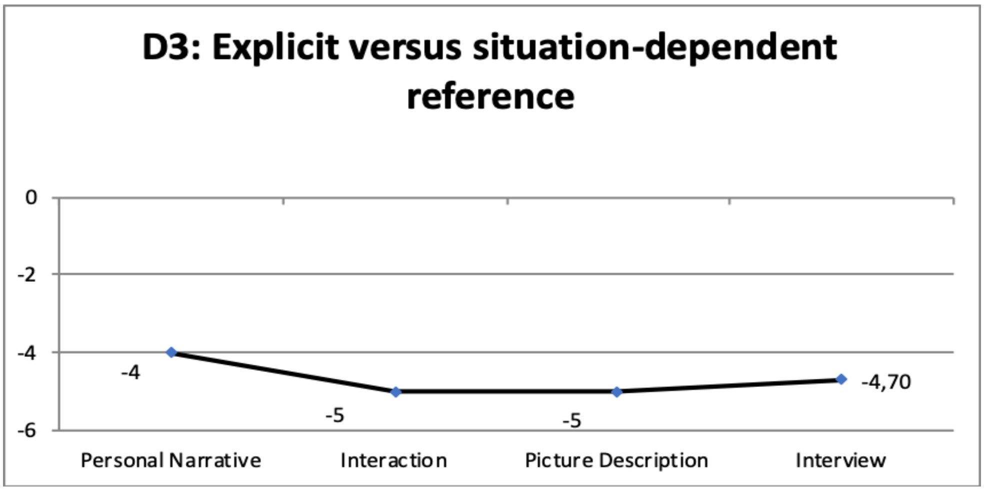

Figure 3. Interview scores on Dimension 3

The score of the Personal Narrative Component on this dimension (-4.5) is closer to face-to-face conversations in Biber (1988) than that of the Description Component (-5), which is closer in turn to telephone conversations. The score difference between these two components (0.5) seems to indicate that our speaking tasks do not play an important factor in the way interview can be linguistically profiled on this particular dimension. This is confirmed by the Duncan's Multiple Range Test for D3, which shows that the three corpus components are not significantly different from each other. Table 8 shows the results of the test.

Table 8. *Effect of speaking task on D3 profiling*

Dimension 3: Explicit versus situation-dependent reference		
Duncan Grouping	Mean	Speaking task
A	-3.9823	Personal Narrative
A	-4.7660	Interaction
A	-4.9726	Picture Description

Alpha 0.05
Error Degrees of Freedom 226
Error Mean Square 5.44739
Harmonic Mean of Cell Sizes 76.20474

Number of Means	2	3
Critical Range	1.255	1.321

The linguistic features which are representative of explicit reference discourse include, in decreasing order of significance, *wh*-relative clauses in object positions, pied piping constructions, *wh*-relative clauses in subject positions, phrasal coordination and nominalizations. Other linguistic features are typically representative of dependent reference discourse: time adverbials, place adverbials and adverbs. Linguistic features that showed negative loadings on this factor such as place and time adverbials, showed frequencies of use unusual in interviews texts in Biber (1988). Table 9 lists the normalized mean of all these linguistic features.

Table 9. *Linguistic features which are representative of the explicit reference dimension*

	object *wh*-relative clauses	subject *wh*-relative clauses	phrasal coordination	nominaliz -ations	place adverbials	time adverbials
Personal Narrative	0.6/1000	1.7/1000	1.8/1000	13.1/1000	12.5/1000	6/1000
Interaction	0.3/1000	0.9/1000	1.7/1000	16.4/1000	12.1/1000	7.1/1000
Picture Description	0.3/1000	1.9/1000	1.9/1000	3.6/1000	6/1000	12.5/1000
Corpus mean	0.4 /1000	1.6 /1000	1.7 /1000	8.1 /1000	10/1000	8.8/1000

Dimension 4: Overt expression of persuasion

This dimension is associated with the expression of own point of view or with the use of argumentation to persuade the interlocutor. The score of the whole interview corpus (-1.02) is far below than that of interview texts (1) and spontaneous speeches (0.3) in Biber (1988). On this dimension, our interview corpus behaves in a similar way to adventure fiction (-1.2) or biographies (-0.7). Figure 4 shows the scores of the components of our corpus on dimension 4.

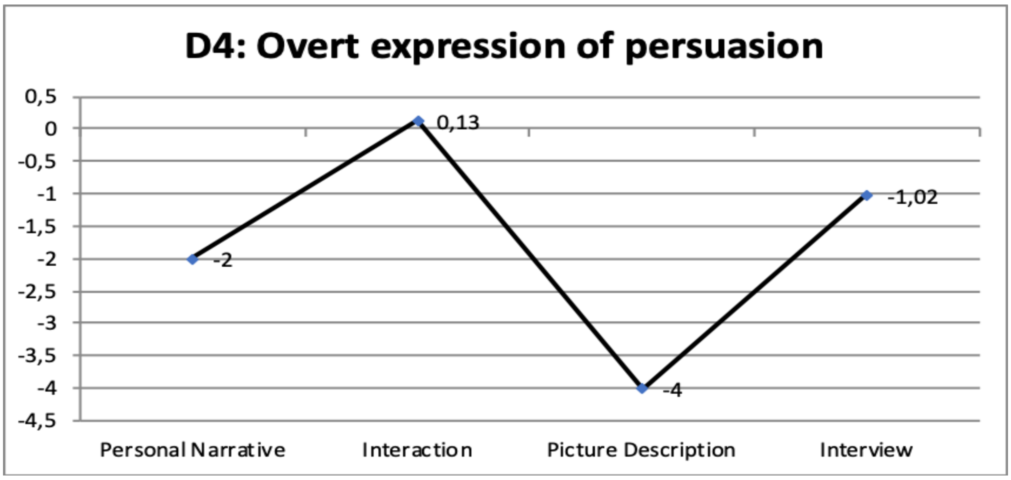

Figure 4. Interview scores on Dimension 4

The score of the Personal Narrative Component on this dimension (-2) is closer to the score of face-to-face conversations (-0.3) in Biber (1988) than the Description Component (-4), which is closer in turn to broadcasts (-4.4). The score difference between the Interaction and the interview Description Components (4.13) seems to indicate that our speaking does play an important factor in the way interviews can be linguistically profiled on this particular dimension. This is confirmed by the Duncan's Multiple Range Test for D4, which shows that all three corpus components are significantly different from each other. Table 10 shows the results of the test.

Table 10. *Effect of speaking task on D4 profiling*

Dimension 4: Overt expression of persuasion		
Duncan Grouping	Mean	Speaking task
A	0.1311	Interaction
B	-1.9848	Personal Narrative
C	-3.8735	Picture Description
Alpha 0.05 Error Degrees of Freedom 226 Error Mean Square 8.902971 Harmonic Mean of Cell Sizes 76.20474 Number of Means 2 3 Critical Range 0.953 1.003		

The linguistic features that are representative of this dimension include, in decreasing order of significance, infinitives, prediction modals, suasive verbs, conditional subordination, necessity modals and split auxiliaries. Table 11 lists the normalized mean of all these linguistic features.

Table 11. *Linguistic features which are representative of the persuasion dimension*

	infinitives	prediction modals	suasive verbs	conditional subordination	necessity modals	split auxiliaries
Personal Narrative	10/1000	4.3/1000	1/1000	2/1000	2.5/1000	2.9/1000
Interaction	15/1000	7.3/1000	1.7/1000	4.1/1000	3.1/1000	2.8/1000
Picture Description	20.1/1000	5.4/1000	1.2/1000	1.1/1000	0.7/1000	0.6/1000
Corpus mean	15.4/1000	5.4 /1000	1.2/1000	2.3/1000	2.4/1000	2.1/1000

Dimension 5: Abstract non-abstract information

This dimension distinguishes discourse with a highly abstract and technical informational focus from discourse which lacks that quality. Academic texts appear at one end of this continuum, while telephone conversations qualify for a type of text where interlocutors share information which is non-abstract and informal (Biber, 1988, p. 113). Figure 5 shows the scores of the components of our corpus on dimension 5.

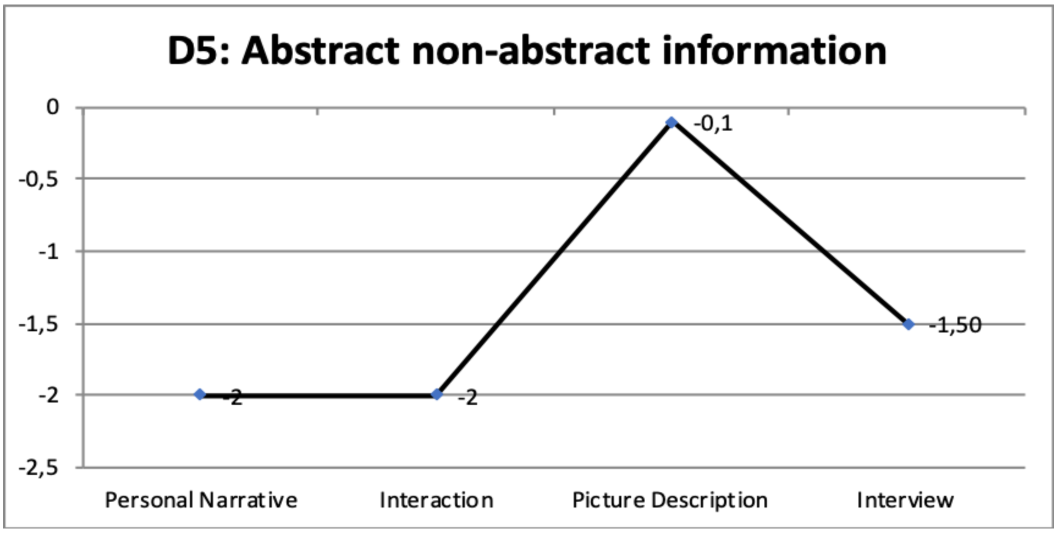

Figure 5. Interview scores on Dimension 5

The mean interview score (-1.5) is close to that of broadcasts (-1.8) and interview texts (-2) in Biber (1988). The Personal Narrative Component score on this dimension (-2) is identical to that of the Interaction Component and to that of the interviews texts in Biber

(1988), while the Picture Description Component (-0.1) overlaps the score of popular lore. The score difference between the Personal Narrative and the Picture Description Components (1.9) seems to indicate that the speaking task does play an important factor in the way interviews can be linguistically profiled on this particular dimension. This is confirmed by the Duncan's Multiple Range Test for D5, which shows that the Picture Description Component is significantly different from the other two tasks. Table 12 shows the results of the test.

Table 12. *Effect of speaking task on D5 profiling*

Dimension 5: Abstract non-abstract information		
Duncan Grouping	Mean	Speaking task
A	-0.1179	Picture Description
B	-1.7865	Personal Narrative
B		
B	-2.3744	Interaction
Alpha 0.05 Error Degrees of Freedom 226 Error Mean Square 16.29063 Harmonic Mean of Cell Sizes 76.20474 Number of Means 2 3 Critical Range 1.288 1.356		

The linguistic features which are representative of this dimension include, in decreasing order of significance, conjuncts, agentless passives, adverbial past participial clauses, *by*-passives, past participial whiz-deletion and other adverbial subordinators (other than cause, concession and condition). Table 13 lists the normalized mean of all these linguistic features.

Table 13. *Linguistic features which are representative of the abstract dimension*

	conjuncts	agentless passives	adverbial ppl. clauses	*by*-passives	ppl. whiz-deletions	other adverbial subordinators
Personal Narrative	1.6/1000	2.7/1000	/1000	0.4/1000	0.5/1000	5.9/1000
Interaction	1.2/1000	1.8/1000	/1000	0.1/1000	0.5/1000	6.8/1000
Picture Description	3.2/1000	7.9/1000	/1000	0.2/1000	0.6/1000	7.5/1000
Corpus mean	2.2/1000	5.2/1000	/1000	0.2/1000	0.5/1000	7/1000

Discussion

This study explores the application of corpus-based methods in LTA and LCR that go beyond test design and validation. By applying MDA to a corpus of 78 interviews with native speakers of English we have tried to provide insights into the nature of spoken tasks from a variationist perspective and, in particular, into the potential of the L2 interview to bring forth linguistic features that would be expected to be characteristic of the spoken register. The adoption of this approach can be key in supporting test validation as conceived by Bachman (1990), who states that "in test validation we are not examining the validity of the test content or of even the test scores themselves, but rather the validity of the way we interpret or use the information gathered through the testing procedure" (Bachman, 1990, p. 238).

Along the lines of Biber and Jamieson (1998, cited in Biber et al., 2004), who found that the linguistic characteristics of the texts in TOEFL exams did not resemble those of the target registers, our results suggest that the different tasks determine the range of linguistic features produced by speakers. For example, the normalized frequency of present tense verbs (118/1000) in the Picture Description Component is considerably higher than in the other two components, almost doubling the frequency of this feature in the Personal Narrative Component (69.9/1000). Could we say then that the picture description task creates the conditions for the use of the simple present tense? What if a speaker adopts a different perspective and decides to tell the painter/young lady story relying on the simple past? Our data exclude this possibility. The range of uses of the present tense in the Picture Description Component goes from 62.1/1000 to 198.4/1000 (SD = 28.7), that is, every speaker in the sample used at least almost the same amount of present tense verbs forms than the mean count for the Personal Narrative Component (69.9/1000).

In the context of learner language assessment, a speaking task is in many ways a speech event where learners are expected to show their competence. If this competence, or level of competence, is matched against the expectations of the examiner/evaluator or against a reference norm, can-do statements, and we all agree that even native speakers' intuitions are not always reliable (Sampson, 2007), it is urgent that we examine how these expectations are shaped by the use of a given register in the community of proficient speakers, i.e. native speakers. One of the types of findings that can be instrumental in this area is that, according to the Duncan Multiple Range Test (see Section 4), the Interaction Component is significantly different from the other two components on Dimension 1 (involved vs. information production), or put another way, the Personal Narrative and the Picture Descriptions Components yield significantly different language.

Biber and Conrad (2010, p. 16) have indicated that "the register perspective characterizes the typical linguistic features of text varieties, and connects those features functionally to the situational context of the variety". This is where corpus linguistics, and particularly MDA, can inform language proficiency evaluators about the complex relationships that govern the use of discrete linguistic features and how texts conform our own understanding of how registers work. Despite the differences between the Interaction Component on the one hand, and the Personal Narrative and the Picture Description components on the other, all three score high on Dimension 1, which profiles them as speaking tasks where speaker's involvement is expected, above interviews or personal letters in the original Biber (1988) study. When examining the linguistic features which are characteristic of more information-oriented registers such as official documents or academic prose, we can see why the interview texts in Biber (1988) are found lower on Dimension 1 than any of the components in our corpus. For example, the normalized count for nouns in the interview register (160.9/1000) is only higher in the Personal Narrative Component (164.7/1000), while prepositions in interviews are more abundant (108/1000) or attributive adjectives (55.3/1000) are infrequent in Personal Narratives (17.2/1000). This shows again how particular speaking tasks are not valid in terms of eliciting certain linguistic features, which calls for a re-examination of the role of interviews and speaking tasks in gathering information about the grammar of learners. In this sense, the Picture Description Component shows very little potential for the use of attributive adjectives (10.1/1000) or, more noticeably, predicative adjectives (2/1000). Consider examples (1) and (2) from our data.

(1)

Speaker: erm well he's drawing her in this picture and then It looks like she doesn't like the way he's drawn her in this one but her the facial expressions

Interviewer: mm

Speaker: erm she like is doesn't like the way she's portrayed she doesn't like the way she looks and he's obviously gone and changed it to make her nicer in the picture obviously to impress friends who look at it like she's been painted nicer something beautiful cos people are gonna look at it and it's her so she wants them to think she looks nice. (CAOS-E C2-3)

(2)

Speaker: okay yeah erm well there's a painter and then there's erm a model who is having here self-portrait done and erm the first picture yeah sets the scenario nd then he says to her to the to the model erm what do you think so far and she doesn't look too pleased and saying that doesn't look anything like me she's unattractive o she she she obviously said well you better do something you better make this better this picture so she does it sits back down and he starts to paint away again and then she looks. she still doesn't look very happy with it in the end and she's saying to some friends or some other people do you think this looks like me. so she's not very happy with the painting okay. (CAOS-E C2-27)

While in (2) we find more adjectives[4], it is apparent that the two speakers do not use much adjectival description, relying more on present tense and the use of nouns to convey the idea of what the situation is about. This tendency is backed up by the corpus data. However, assessment of learners' lexis is commonly found in rubrics. Some researchers have seen that the correct use of adjective order (Lightbown & Spada 1990) or native-like intensification of adjectives (Lorenz 1999) are good indicators of language proficiency. Notwithstanding, other than in interviews which make use of cues to elicit discrete language features, it is extremely difficult to determine which registers or speaking tasks are more likely to yield which linguistic features.

As for Dimension 2, narrative versus non-narrative concerns, the Duncan's Multiple Range Tests corroborates that the three components of our corpus differ significantly from each other, which shows that the three speaking tasks offer distinct profiles on this dimension. While the Interaction Component overlaps with the mean score of telephone conversations (-2) in Biber (1988), the Personal Narrative Component overlaps with the mean score of face-to-face conversations (-0.7). The Interaction Component showed the lowest mean on this dimension, qualifying as the least narrative sub-register in our corpus data. The normalized frequency count of 3rd person pronouns (12.3/1000) and public verbs (2.2/1000) is the lowest in our corpus. By contrast, one of the reasons which may account for the high frequency of 3rd person pronouns in the Picture Description Component (94.5/1000) is the nature of the story going on in them, which includes the elaboration on a sequence of pictures involving a painter and a young lady being portrayed. This clearly favours the use of anaphoric reference and, together with the constraints on online processing in spoken discourse, created the conditions for this comparatively higher frequency of 3rd person pronouns. (3) is another example sample from our research corpus.

(3)

Speaker: okay there's this this woman has gone to the to an artist for a portrait he does the portrait which is a true representation of her and she doesn't like it

Interviewer: mhm

Speaker: she wants to be made more beautiful than she thinks she is so he gets she gets him to redo it and shows off the portrait to her friends showing her as an nice attractive young woman clearly she isn't sadly so she she wants the portrait to give her a picture of what she sees herself as

Interviewer: mhm

Speaker: rather than what the world sees her as (LOCNEC-15)

All three speaking tasks scored low on Dimension 3, explicit versus situation-dependent reference, finding themselves between the ranges of face-to-face conversations (-4) and telephone conversations (-5.2). The tasks that were used to elicit spoken language proved to have no discriminatory power for this dimension of use, which was corroborated by the

[4] The adjective variation index is 0.12 for the first simple and 1.4 for the second.

Duncan’s Multiple Ranges Tests. The fact that the frequency distribution of phrasal coordination and *wh*-relatives in object and subject positions is similar in the three corpus components, and that their scores overlap registers such as face-to-face conversations or telephone conversations in Biber (1988), which are neither explicit nor heavily situation-dependent registers, seem to indicate the lack of adequacy of these registers or elicitation tasks for the assessment of learner language along the functional underpinnings of D3. This finding is supported by Biber, Reppen and Conrad (2002, p. 46), who stated that there is “comparatively little linguistic variation among spoken registers, apparently because they are all constrained by real-time production circumstances”. The same applies to Dimension 5, abstract versus non-abstract information, where significant differences where only found between the Picture Description Component, on the one side, and the Narrative and the Interaction components on the other, which in actual fact yielded the same score on this dimension. As one may expect, the three components showed very little power to generate abstract language of the type found in academic prose. Despite the Personal Narrative Component, it seems that the restrictions imposed by spoken communication were stronger than the thematic orientation of interviews for this component, where speakers were invited to talk about a book, a film or a journey that had influenced their lives. Another alternative explanation may be that the involvement dimension actually was favored by the speakers, defying the restrictions imposed by the university setting where the interviews took place.

However, what has been discussed about Dimensions 3 and 5, does not apply in the case of Dimension 4, overt expression of persuasion, where all three components were profiled in a significant different way. This finding may be of interest to EFL educators and test writers as the expression of one’s point of view is a pivotal communicative function across the foreign language learning curriculum, from beginner to advanced levels. Contrary to the situation on Dimension 3, there is a huge difference between the mean scores of the Interaction Component (0.13) and the Picture Description Component (-4). Clearly, this last speaking task yields fewer opportunities for the expression of one’s own point of view. The Interaction Component score is closer to NS registers such as spontaneous speeches (0.3) and face-to-face conversations (-0.3) than the Personal Narrative Component (-2) and the Picture Description (-4) Components. The frequency of prediction and necessity modals, suasive verbs as well as conditional subordination is much higher in the Interaction Component, which explains its power to generate communication where persuasion and point of view are evaluated.

Our research methodology provides usage evidence of NS language in registers that have not traditionally been included in major reference corpora such as the Brown Corpus or the BNC. In contrast, the interview corpus used is defined by the speaking tasks used when collecting learner language data. Principled corpora are made up of registers that represent NS use of the language, such as face-to-face conversations, sermons, radio broadcasts or fiction. It is interesting that these representative corpora have never included speaking tasks that are ironically so pervasive in language assessment and, accordingly, in language education. This fact has prevented learner language researchers from establishing more robust comparability analyses between NS and NNS language, at least in spoken communication.

The type of findings we have discussed in our paper is in keeping with the claims of researchers in the field of corpus linguistics (Flowerdew, 2009) which call for the inclusion of contextualization in corpora. Furthermore, our research addresses concerns expressed long ago regarding validity in general and content relevance (or validity) in particular such as those to which Bachman (1990) drew attention: “the problem with language tests, of course, is that we seldom have a domain definition that clearly and unambiguously identifies the set of language use tasks from which possible test tasks can be sampled, so that demonstrating either content relevance or content coverage is difficult” (Bachman, 1990, p. 245).

Based on the data we have discussed in this paper, the LINDSEI-format interview can be considered a complex register on its own, with peculiarities which bring it closer to conversational language on most dimensions of use; but also a complex register which is very sensitive to the tasks which are selected to elicit language. On Dimensions 2 and 4 all three components differed from each other in a significant way, while on Dimensions 1 and 5 only the Interaction and the Description Components, respectively, behaved differently. Further research should examine each of these speaking tasks more closely so as to determine the potential benefits and drawbacks for language assessment and learner language research in the context of register and language variation.

Conclusions

The results of our MDA of native speaker language suggest that L2 interviews can be instrumental in creating the context for a more complex assessment of learner language proficiency, as the different speaking tasks involved have the potential to yield sub-registers of different nature. By exploring the characteristics of different speaking tasks, we have shown practical ways in which new registers can be linguistically profiled. This profiling is of interest in areas such as language assessment, where language interviews are widely used to evaluate the speakers’ communicative competence, but also in the field of learner language research, where corpora such as LINDSEI or the TLC will unlock new perspectives on learners’ spoken communication in similar ways as the *International Corpus of Learner English* (ICLE; Granger et al., 2009) did for the written mode.

Despite the limitations of our study, namely the number of interviews included and the exclusive use of the British variety of English, our research sheds light on central issues which affect language assessment and learner language research methodology. Moreover, the fact that studies like this are still very few in number (mainly Biber and Jamieson, 1998 and Biber et al., 2004) limits our capacity to relate our findings to previous work carried out along the same lines. These three limitations provide evidence that the potential of MDA of NS data to inform LTA is still under-exploited, which, on the other hand, hopefully opens up new ways to future work.

Further analyses of each of the speaking tasks of our corpus will contribute to unveil the interplay between linguistic features, the functional dimensions of use in the MDA tradition and the role of these features in the assessment of language proficiency in spoken communication.

Disclosure Statement

No potential conflict of interest was reported by the authors.

References

American Council on the Teaching of Foreign Languages. (1999). *ACTFL Proficiency Guidelines –Speaking*. Retrieved from http://www.actfl.org/sites/default/files/pdfs/public/Guidelinesspeak.pdf

Aguado, P., Pérez-Paredes, P. & Sánchez, P. (2012). Exploring the use of multidimensional analysis of learner language to promote register awareness. *System*, *40*(1), 90-103.

Alderson, J. C. (1996). Do corpora have a role in language assessment? In *Using Corpora for Language Research.* J.A. Thomas and M.H. Short (eds.), 248-259. London: Longman.

Barker, F. (2010). How can corpora be used in language testing? In *The Routledge Handbook of Corpus Linguistics.* A. O'Keeffe & M. McCarthy (eds.), 633-645. Abingdon, UK: Routledge.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Biber, D. (1988). *Variation across Speech and Writing.* Cambridge: Cambridge University Press.

Biber, D. (2003). Variation among university spoken and written registers: A new multi-dimensional approach. In *Corpus Analysis. Language Structure and Language Use*. P. Leistyna & C. F. Meyer (eds.), 47-70. Amsterdam & New York: Rodopi.

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

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V., Csomay, E., & Urzua, A. (2004). Representing language use in the university: analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus, Report Number: RM-04-03, Supplemental Report Number: TOEFL-MS-25, Educational Testing Service, Princeton, NJ.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E. (1999). *Longman Grammar of Spoken and Written English.* Essex: Pearson Education.

Biber, D., Reppen, R. & Conrad, S. (2002). Developing linguistic literacy: Perspectives from corpus linguistics and multi-dimensional analysis. *Journal of Child Language* 29 (2), 458 - 462.

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

Connor-Linton, J. & Shohamy, E. (2001). Register validation, oral proficiency, sampling and the promise of multi dimensional analysis. In *Variations in English*. S. Conrad, & D. Biber (eds.), 124-137. Harlow: Pearson Education.

Conrad, S. (2001). Variation among disciplinary texts: A comparison of textbooks and journal articles in biology and history. In *Multi-dimensional studies of register variation in English*. S. Conrad & D. Biber (eds.), 94-107. Harlow, England: Pearson Education.

Conrad, S. & Biber, D. (2001). *Variation in English. Multi-Dimensional Studies*. Harlow: Pearson Education.

De Cock, S. (1998). Corpora of learner speech and writing and ELT. In *Proceedings of the International Conference on Germanic and Baltic Linguistic Studies and Translation.* A. Usoniene (ed.), 56-66. Vilnius: Homo Liber.

De Cock, S. (2004). Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures 2*, 225-246.

Díez-Bedmar, M.B. (2018). Fine-tuning descriptors for CEFR B1 level: insights from learner corpora, *ELT Journal*, *72*(2), 199-209.

Ferrara, S. (2008). Design and psychometric considerations for assessments of speaking proficiency: The English Language Development Assessment (ELDA) as illustration images. *Educational Assessment*, 13(2), 132-169.

Flowerdew, L. (2009). Applying corpus linguistics to pedagogy: A critical evaluation. *International Journal of Corpus Linguistics*, *14*, 393–417.

Gablasova, D., Brezina, V. & McEnery, T. (2019). The Trinity Lancaster Corpus. Development, description and application. *International Journal of Learner Corpus Research* 5(2), 126–158.

Gilquin, G., De Cock, S., & Granger, S. (2010). *The Louvain International Database of Spoken English Interlanguage. Handbook and CD-ROM*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Gilquin, G. & Gries, S. T. (2009). Corpora and experimental methods: a state-of-the-art review. *Corpus Linguistics and Linguistic Theory*, *5*(1), 1-26.

Iwashita, N., Brown, A., McNamara, T. & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*(1), 24-49.

Johnson, M. (2001). *The Art of Nonconversation: A Reexamination of the Validity of the Oral Proficiency Interview.* New Haven: Yale University Press.

Kunnan, A. J. (1998). Approaches to validation in language assessment. In A. J. Kunnan (Ed.), *Validation in language assessment*, pp. 1-16. Mahwah, N.J.: LEA.

Lightbown, P. & Spada, N. (1990). Focus on form and corrective feedback in communicative language teaching: Effects on second language learning. *Studies in Second Language Acquisition* 12, 429-448.

Lorenz, G. (1999). *Adjective intensification - learners versus native speakers. A corpus study of argumentative writing*. Language and Computers: Studies in Practical Linguistics 27. Amsterdam & Atlanta: Rodopi.

McNamara, T.F., Hill, K. & May, L. (2002). Discourse and assessment. *Annual Review of Applied Linguistics, 22*, 221–43.

Neary-Sundquist, C. A. (2009). *The role of task type and proficiency level in second language speech production.* PhD Dissertation, Purdue University.

Pérez-Paredes, P., & Bueno-Alastuey, M. (2019). A corpus-driven analysis of certainty stance adverbs: Obviously, really and actually in spoken native and learner English. *Journal of Pragmatics*, 140, 22-32.

Ricardo-Osorio, J. G. (2008). A study of foreign language learning outcomes assessment in U.S undergraduate education. *Foreign Language Annals, 41*(4), 590-610.

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing*, *11*(2), 99-123.

Shohamy, E., Donitsa-Schmidt, S. & Waizer, R. (1993). The effect of the elicitation mode on the language samples obtained in oral tests. Paper presented at the 15th Language Testing Research Colloquium, Cambridge, England.

Taylor, L. and Barker, F. (2008). Using Corpora for Language Assessment. In *Encyclopedia of Language and Education, second edition,* E. Shohamy & N. H. Hornberger (eds.), vol. 7, Language Testing and Assessment, 241-54. New York: Springer.

Van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: oral proficiency interviews as conversation. *TESOL Quarterly, 23*, 489-508.

Copyrights