METHOD ARTICLE

An Approach to Assessing the Linguistic Difficulty of Tasks

Gabriele Pallotti

This article proposes an approach to assessing the linguistic difficulty of tasks, that is, the linguistic features involved in performing a communicative task that may make it more or less challenging for language learners. The procedure follows the methodology proposed by Pallotti (2019) for operationalizing task interactional difficulty. This consists, firstly, in identifying what linguistic-communicative features are particularly difficult for language learners, based on previous research showing that they appear late in the course of acquisition. Secondly, native speakers' performance is observed in order to determine which tasks most involve these difficult linguistic features. The dimensions observed in this study concern lexical diversity and sophistication, morphological complexity, and length and depth of syntactic constructions. Data come from 10 native speakers of Italian performing 5 communicative tasks. Results show that different dimensions of linguistic difficulty are relatively independent of each other, and that inter-individual variation is rather limited as regards the lexicon and morphology, but more pronounced for syntax. Implications for SLA research, Task-Based Language Teaching and Task-Based Language Assessment are discussed.

1. Introduction

Over the last decades a considerable body of research has accumulated on the relationship between the characteristics of communicative tasks and their effects on second language performance. The results achieved to date, however, are not very clear and consistent, and this is frequently attributed to the fact that a large number of measures and operationalizations have been proposed, with little attention to construct validity and the replicability of results (Ellis, 2018; Long, 2015; Plonsky & Kim, 2016).

In recent years several works have appeared with the aim of clarifying key constructs, both with regards to the dependent variables of complexity, accuracy and fluency and the independent variable of task complexity, or difficulty (in this article the term difficulty will be preferred, for reasons that will be explained in the next section) (Bulté & Housen, 2012; Norris & Ortega, 2009; Pallotti, 2009; Révész, 2014; Révész, Michel et al., 2016; Sasayama, 2016). Continuing along this line, this article presents an approach to explicitly defining and operationalizing linguistic difficulty, one of the aspects that makes a task more or less challenging.

The argument follows the approach proposed by Pallotti (2019) to assess task interactional difficulty, extending it to a new domain, i.e. linguistic difficulty. First, some linguistic features will be deemed to be more difficult than others, based on previous research showing that they systematically appear later in interlanguage development. Indeed, "a language feature is more difficult than another if its processing and learning requires more time and/or more mental activity" (Housen & Simoens, 2016, p. 166). Then, the performance of native speakers on five tasks will be analysed, in order to assess whether different tasks elicit variable amounts of difficult features. These results will be used to establish, in an empirically grounded, explicit way, whether one task is more difficult than another from a linguistic point of view. The relative difficulty of the same tasks may change with respect to other dimensions, such as interactional difficulty, reasoning demands or pragmatic constraints. The point made here is that different dimensions of task difficulty can and should be assessed independently, in order to arrive at a clearer picture of the demands that different tasks make on task performers and, as a consequence, a better understanding of how these demands impact on participants' communicative behaviours.

Linguistic difficulty is a key element contributing to a task's global difficulty, and it is mentioned in virtually all accounts of L2 communicative tasks, beginning with

University of Modena and Reggio Emilia, IT
gabriele.pallotti@unimore.it

Candlin's (1987) seminal paper, where it was named "code complexity", a term subsequently adopted by Skehan (1992, 1998). Linguistic difficulty is also relevant to Task-Based Language Teaching (TBLT), as it is one of the criteria that may inform a task-based syllabus (Baralt, Gilabert & Robinson 2014), and in language testing and assessment, where the linguistic difficulty of different tasks needs to be graded according to proficiency and controlled for in multiple editions of the same test (Elder et al., 2002).

2. Literature review

The literature on how task characteristics impact on linguistic-communicative performance is vast and constantly growing (see recent reviews by Ellis, 2018; Long, 2015; Wen & Ahmadian, 2019). This section will not attempt to provide a comprehensive synthesis but will rather focus on some areas that are most relevant for the present contribution. Firstly, a terminological discussion will argue for the use of the expression 'task difficulty' instead of 'task complexity', if what is meant are the demands that a task makes on its performers. This will be followed by a review of previous definitions of linguistic difficulty and of the studies that allow one to empirically assess it according to a commonly recognized criterion, that is, late emergence in the course of L2 acquisition. Finally, the use of native speakers' baseline data for assessing task demands will be scrutinized, as it has not been very common in the past but is becoming more widespread and is also advocated in this contribution.
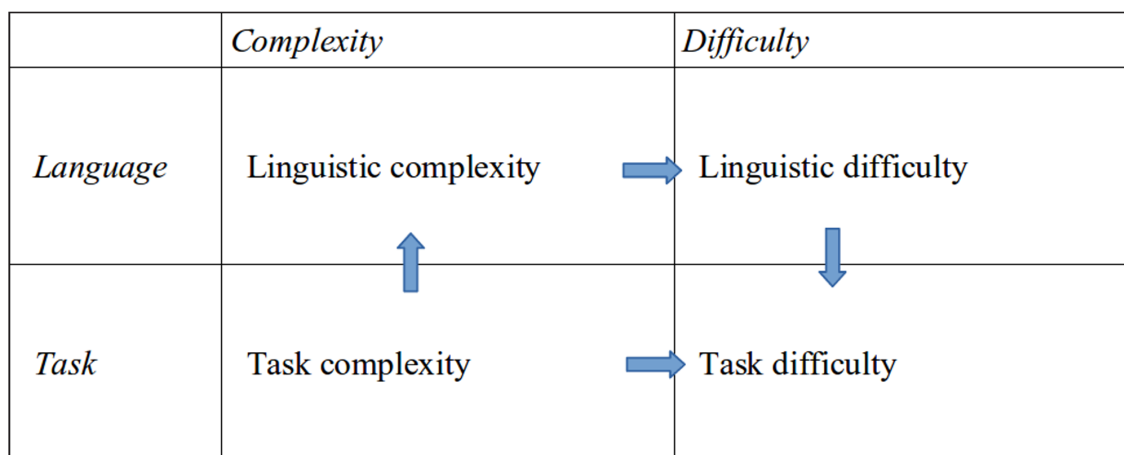
2.1. Complexity and difficulty

In this article the expression *task difficulty* will be preferred to *task complexity*, which has been prevalent in the SLA literature over the past two decades or so; this terminological choice thus needs some justification.

As a matter of fact, most early works referred to task "difficulty" (e.g. Brindley, 1987; Candlin 1987, Nunan, 1989; Skehan 1992). In the language testing literature, too, the term difficulty is almost exclusively employed (e.g. Elder et al., 2002; Fulcher & Márquez Reiter, 2003). One of the first to consistently use the expression "task complexity" was Peter Robinson (1995, 2001), after which the term gained more and more ground in SLA research.

This terminological choice, however, is not without problems, mainly because of the polysemy of the term *complexity*, which can mean both an object's structural properties (the number of its parts and of the relations among them) and the cognitive demands faced by human beings when interacting with that object.[1] In the interest of terminological clarity, some authors have proposed that the two notions should be labeled with different terms, such as complexity for the former, and difficulty for the latter (Bulté & Housen, 2012; Housen, in press; Housen & Simoens, 2016; Pallotti, 2009, 2015; Skehan, 2015), which, among other things, would also facilitate research on the relationships between them. This holds for both tasks and linguistic features, that can be said to be more or less complex (composed of several elements with intricate structural relationships) or difficult (posing higher demands on the users). It is certainly possible to study whether and to what extent more structurally complex objects are more difficult for human beings to deal with. However, this is not a reason for using the same term for the cause (structural complexity) and the effect (cognitive difficulty), but actually suggests that the two notions should receive different labels.

Figure 1 graphically depicts the relationships among these constructs. The first column concerns complexity, defined by Rescher (1998, p. 1) as "the number and variety of an item's constituent elements and of the elaborateness of their interrelational structure". Linguistic features or texts may be complex because they contain many different elements (e.g. a high variety of lexical items or morphological processes) or because their relationships are intricate (e.g. long syntactic structures with deeply embedded constituents) (Bulté & Housen, 2012; Pallotti, 2015). This structural linguistic complexity may contribute to linguistic difficulty, that is, to the effort required of a human being to process and master such structures or produce texts containing them (DeKeyser, 2005; Housen, in press; Housen & Simoens, 2016; Spada & Tomita, 2010). Linguistic difficulty in turn contributes to task difficulty when a task, in order to be adequately performed, requires many difficult linguistic features. However, this is just one source of task difficulty, which may also be increased by the structural complexity of

	Complexity	*Difficulty*
Language	Linguistic complexity	➡ Linguistic difficulty
Task	Task complexity	➡ Task difficulty

Figure 1: Complexity and difficulty in language and tasks.

the task itself, for instance when the task contains many elements related to another in a variety of ways, or with constraints on their co-occurrence (Skehan, 1998, 2015; Robinson, 2001, 2011, 2015). Finally, according to some theoretical models (e.g. Robinson, 2011), task complexity itself may also lead to the production of more complex linguistic structures and thus contribute, in a more indirect way, to task difficulty.

The arrows in **Figure 1** should not be taken to imply that relationships are circular, as if everything caused everything. There is a clear directionality between complexity and difficulty: as Rescher puts it, "cognitive difficulty reflects rather than creates complexity" (1998, p. 17) or, with specific regard to second language acquisition, "structural complexity can contribute to psycholinguistic complexity or difficulty, but does not coincide with it" (Housen, in press, p. 2). As a matter of fact, the bottom right cell, task difficulty, has arrows pointing to it, but none pointing from it, which means that task difficulty is the (more or less direct) product of many factors, but not their cause.

2.2. Defining and assessing linguistic difficulty in SLA
Notions like "code complexity" (Candlin, 1987; Skehan, 1992, 1998) express the intuition, shared by researchers, teachers and lay people, that some communicative tasks are more difficult than others because they require more complex linguistic structures – such as a varied and sophisticated vocabulary or the use of intricate syntactic and textual structures – and that this complexity leads to higher difficulty for task performers.

These intuitions have been developed in subsequent research, and several criteria have been proposed to establish whether linguistic features are more or less difficult (see reviews by Collins et al., 2009, DeKeyser, 2005, Housen & Simoens, 2016, Housen, in press). Structural complexity is often cited as one of the causes of linguistic difficulty, together with frequency and saliency in the input. Acquisitional timing, on the other hand, is considered to be an effect of linguistic difficulty: A structure may be said to be more difficult if it is acquired late, that is, if it appears at relatively advanced levels of L2 development.

Based on these general criteria for establishing linguistic difficulty, the following aspects may be examined in order to identify more specific constructs and their measures. The list does not exhaust all the features that have been shown to develop over time in L2 acquisition, but selects only some, chosen among those most investigated in previous research and that are not limited to particular languages.

2.2.1. Lexicon
Several studies have shown that in initial interlanguage varieties the lexicon tends to be repetitive and mostly contains high frequency words; rarer words, which can also be called more sophisticated, are acquired later, as well as the ability to use a varied lexicon, i.e. a high proportion of lexical types compared to the tokens produced (De Clercq, 2015; Dóczi & Kormos, 2016; Kang, 2013; Treffers-Daller, 2013; Yu, 2010). Therefore, a task

requiring a varied lexicon (higher structural complexity) with several low-frequency words (higher acquisition difficulty) will be considered to be more difficult than one implying just a small set of frequent words.

2.2.2. Morphology
One of the first and most replicated findings of SLA research is that inflectional morphology is absent or very limited in basic interlanguage varieties, as it develops later, with variable speed and outcomes depending on individual factors and on the structural complexity of the system to be acquired (for recent contributions and overviews of previous literature, Brezina & Pallotti, 2019; De Clercq & Housen, 2019). For these reasons, a task involving the use of a wide range of morphological processes can be said to require greater skills, and thus be more difficult, than one involving just a few morphological processes. The range of morphological processes can be calculated using the Morphological Complexity Index (Pallotti, 2015; Brezina & Pallotti, 2019), which measures the variety of morphological types appearing in a text.

2.2.3. Syntax
Languages also differ as regards syntax, with some having just one or two basic word orders, and others displaying a wide array of constructions with several constraints on their occurrence, based for instance on the type of constituency relation or illocutionary force. Thus, a task involving certain linguistic constructions or speech acts can be easy in one language and difficult in another. Nevertheless, research shows that, in general, the initial phases of second language acquisition are characterized by syntactically simple constructions, i.e. short and relatively independent of each other; only later are learners able to control more far-reaching structures, consisting of a large number of words or clauses. Vercellotti (2018, p. 7), in her study of the longitudinal development of L2 English speech, provides the following examples: *Next time I can pay them back* (less complex); *if I don't like this man and I don't want to have a next date I think they pay the bill first* (more complex), and shows that more structurally complex constructions tend to increase over time.

Measures such as mean length of production unit, number of clauses per unit and subordination ratio all represent this greater complexity of syntactic structures, and they have been shown to steadily increase at least from initial to intermediate levels, while at more advanced levels there is stabilization with greater variability, probably linked to individual stylistic preferences (for recent contributions and overviews of previous literature, De Clercq & Housen, 2017; Kuiken et al., 2019; Vercellotti, 2018; for Italian, Chini, 2003). It can thus be maintained that tasks involving the production of long and complex syntactic structures, containing several elements linked together, require more skills and are therefore more difficult from a linguistic point of view.

2.3. Native speakers' task performance
After having established, on the basis of empirical research, which linguistic features are more difficult as they take

longer to be acquired, it is necessary to observe which tasks most require these features. Since we are concerned with difficulty for additional language users, it would seem natural to observe their performance. However, this is more problematic than it seems. In fact, if these learners were not to produce difficult linguistic behaviours in a task, it would be impossible to say whether this is due to the fact that the task does not require them, or to the fact that their skills do not allow it. Previous research has in fact shown that L2 proficiency systematically mediates between task characteristics and linguistic performance (e.g. Malicka & Levkina, 2012; Sasayama, 2016).

To overcome this problem, one may look at the performance of native speakers (Ellis, 2011; Long, 2015, p. 239; Pallotti, 2019), who form a more homogeneous group than learners, at least as regards the fundamental structures of "basic language cognition" (Hulstijn, 2015, 2019). As far as this type of language structures is concerned, native speakers consistently reach the highest scores, representing a sort of ceiling with respect to the wider range of scores obtained by learners at different levels (Abrahamsson & Hyltenstam, 2009; Granena & Long, 2013). Of course, it may be possible for some non-native speakers to reach the same levels as the natives, at least in some areas, so that the whole category may be labelled, in more general terms, "top language performers", to refer to individuals whose performance is at or close to ceiling levels. In any case, observing which structures are used by these top language performers in different tasks provides an indication of how the tasks themselves, rather than the speakers' (in)capacities, favour or limit their use. In other words, the observation of top language performers, who have at their disposal the whole range of structures, from the easiest to the most difficult, makes it possible to more directly observe how different tasks involve the use of more or less difficult structures.

Some previous studies have looked at native speakers' performance on tasks, with the aim of comparing it with that of language learners. For instance, Skehan (2009) observed that native and non-native speakers behaved rather similarly as regards their use of infrequent words and of varied lexicon in personal information exchange and decision-making tasks, while differences were more noticeable in picture-story retellings. For both groups, the two measures varied independently of each other, thus demonstrating that lexical variety and sophistication are independent constructs. Foster and Tavakoli (2009) showed that native speakers' syntactic complexity varied across different narrative tasks depending on storyline complexity, while Ellis (2011) found that syntactic and lexical complexity were different in different types of tasks (reporting a car accident vs giving directions on a map), although manipulating each type of task in order to make it more or less cognitively demanding did not lead to any changes in native speakers' linguistic performance.

There is thus evidence showing that native speakers' linguistic behaviours do indeed vary – like those of second language learners, though sometimes in different ways – depending on task conditions. However, none of these studies saw these variations in native speakers'

performance as indexing higher or lower levels of potential difficulty for language learners, which is the focus that will be taken in this article.

3. The study

Data for this study come from the VIP (*Variabilità dell'Interlingua Parlata* [Variability of Spoken Interlanguage]; Pallotti et al., 2011) corpus, also used by Pallotti (2019). Participants were girls aged 15–20 at the beginning of data collection, attending high schools in Northern Italy. 14 were non-native speakers with a variety of L1s, while 10 were native Italian speakers – this study will look at these only (mean age = 18.0). The relatively small sample used in this study implies that quantitative analyses should be taken as illustrating how the procedure may be practically implemented and indicating areas worthy of further investigation, rather than as making inferential claims about the generalizability of results for this particular set of tasks and participants.

Participants performed a variety of oral communicative activities, so that their linguistic skills could be assessed in a range of contexts. The procedure consisted of two sessions on two different days. The first session involved a series of essentially monologic tasks and began with a semi-structured interview, followed by retelling a silent film and a picture story, then by a map task with the adult interviewer. The second session proposed more interactive tasks, with participants working in pairs. There was another map task, this time with the peer, and two information-seeking activities, one requiring them to plan a school trip, the other to select a present for a friend. Both these tasks involved making a number of phone calls to shops, travel agencies, restaurants and hotels, and to a list of "experts" (both youths and older adults) who were asked to provide advice and information. Apart from the initial ice-breaking conversation, all the other tasks were presented in a counter-balanced order in different sessions.

Tasks for this project were devised so that they would vary mostly on pragmatic and sociolinguistic dimensions, such as the type of communicative moves to be performed (e.g. initiating, responding, negotiating), monologic vs dialogic activities, social distance between interlocutors (acquaintances vs strangers, peers vs adults). No task manipulations were envisaged to target specific linguistic dimensions, so that all tasks were assumed to involve rather ordinary everyday language of comparable difficulty, a point that should be borne in mind when interpreting the results presented in the next pages.

In this article we will look at native speakers' data from the interview, film retelling, map task with a peer, and from phone calls and face-to-face negotiations during the school trip organization (total corpus size: 71,500 words). Given that the interviews and the school trip organization task lasted much longer than other activities, only the first ten minutes of the first task and the last ten of the second will be analysed here (the choice is due to the fact that these parts of the activities were more uniform across dyads, so that in the interpretation of results intra-task variability would have a lower impact than inter-task

systematic variation). Transcription followed a modified version of the Chat-CA system.

Transcribed data were prepared for quantitative analysis by first dividing them into AS-units (Analysis-of-Speech Unit, Foster et al., 2000) and clauses. This segmentation was carried out by students and research assistants on about 60% of the data, and then checked by the principal investigator; after some initial training and discussions, inter-rater agreement was always over 85%. The remainder of the data were coded by the principal investigator only.

Morphological and lexical analyses were conducted using automatic tools, which implied standardizing orthography and removing from the original transcription all non-verbal behaviour markers, such as pauses, breaths, laughter. Given that Italian is a highly inflected language, lexical diversity and sophistication were calculated on lemmas, obtained with the part of-speech analyser Tree-tagger (Schmid, 1994) and subsequent manual revision.

4. Results

4.1. Lexicon

Lexical variety was assessed with the Moving Average Type-Token Ratio (MATTR, Covington & McFall, 2010), that is, the average type-token ratio (TTR) in fixed-length samples taken from a text (in this case, 250 words, which was slightly less than the shortest text in the corpus). MATTR is calculated by averaging the TTR of multiple text samples, one after another, so that each sample includes all the words of the previous sample except the first, plus a new word, until the end of the text is reached.

Lexical sophistication was calculated as the proportion of words not belonging to the most frequent 2,000 lexical types, deemed the "fundamental" lexicon of Italian (De Mauro, 2016) and computable with the online tool Dylan Text Tools 2.1.9 (www.ilc.cnr.it/dylanlab/apps/texttools). The 2,000 most frequent words list is considered to be an important threshold for lexical richness according to Laufer's (1995) notion of "Beyond 2000".

As shown in **Table 1** and **Figure 2**, lexical diversity values, calculated using the Moving Average Type/Token ratio with a window of 250 tokens (MATTR-250), did

not exhibit large differences across tasks, except for the map task, where a smaller range of types was used (MATTR-250 = 0.36).

Another indicator for a task's lexical difficulty is the proportion of non-basic words used in its performance. This was operationalized as the percentage of words in the text not belonging to the 2,000 most frequent lexemes in Italian. In this domain, too, values do not change very much across tasks, with negotiations and film retelling having the lowest proportion of non-common words (**Table 2** and **Figure 3**).

It is worth noting that the two measures of lexical difficulty, viz. type/token variety and low frequency lemmas, do not go exactly hand-in-hand. For instance, negotiations had the highest lexical diversity, but the lowest proportion of infrequent words; on the other hand, the lexicon for performing the map task was not very basic, but it was rather repetitive.

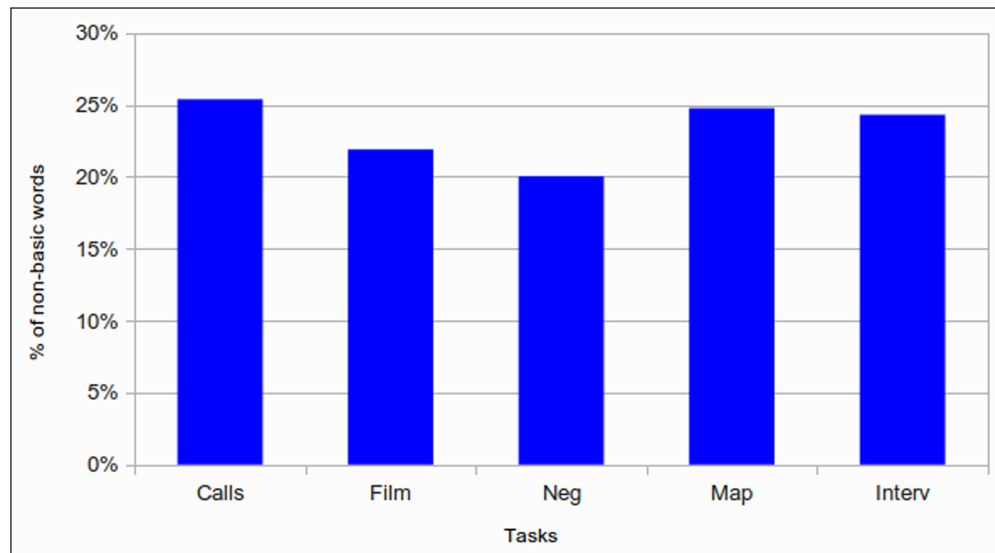**Table 1:** Moving Average Type/Token Ratio (MATTR-250) across tasks.

Task	Mean	SD
Calls	0.44	0.05
Film	0.45	0.03
Neg	0.49	0.04
Map	0.36	0.02
Interv	0.46	0.03

Table 2: Percentage of non-basic words across tasks.

Task	Mean	SD
Calls	25.38%	0.49
Film	21.90%	0.83
Neg	20.03%	0.44
Map	24.77%	2.72
Interv	24.31%	1.72
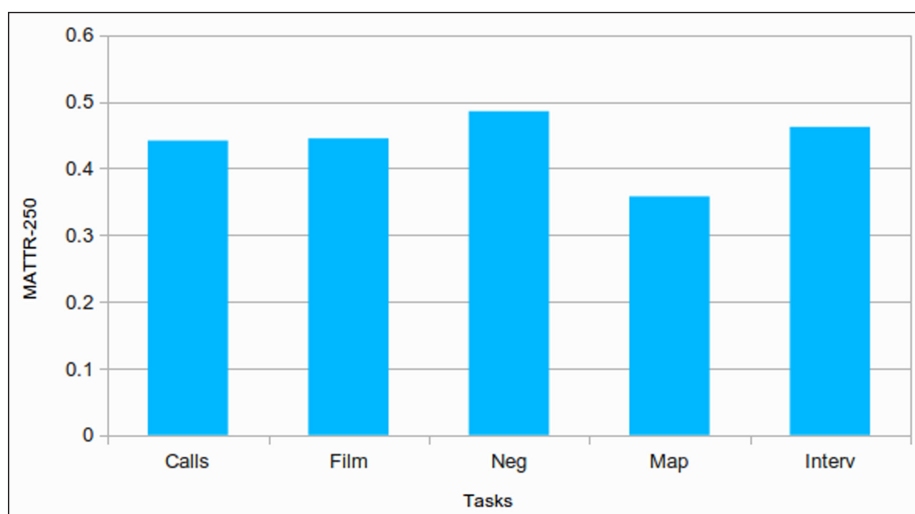
Figure 2: Moving Average Type/Token Ratio (MATTR-250) across tasks.

Figure 3: Percentage of non-basic words across tasks.

4.2. Morphology

The Morphological Complexity Index (MCI, Pallotti, 2015) was computed with the online Morpho Complexity Tool (Brezina & Pallotti, 2015; corpora.lancs.ac.uk/vocab/analyse_morph.php), and was calculated as the average within- and across-sets diversity of samples of 10 verbal exponents, randomly sampled 100 times from each text. This measure thus gives an indication of the variety of verbal inflections used in different tasks.

Table 3 and **Figure 4** show that the MCI values are quite similar for phone calls, negotiations and interviews, which display relatively high scores, all over 12. The variety of verbal exponents was slightly lower in the film retelling (11.24), where the plot was typically told using a few persons of the present tense, and much lower (8.54) in the map task, where most verb forms were in the second person singular of the imperative or in the third person singular of the present tense.

4.3. Syntax

Among the many measures that have been proposed to assess syntactic development in an additional language, two were selected for this study. The first is the mean length of AS-Unit, defined as a main clause or sub-clausal unit with all the dependent clauses attached to it (Foster et al., 2000). This measure provides a general indication of the breadth of unitary syntactic structures. The second measure is the number of dependent clauses per AS-Unit, which more specifically represents the degree of syntactic embedding. Both measures have been extensively applied in the SLA literature on several languages and have been shown to increase at higher proficiency levels in L2 oral productions (e.g. De Clercq & Housen, 2017; Vercellotti, 2018).

Results for the syntactic analysis are more variegated than for other linguistic dimensions, with rather conspicuous variations across tasks. As regards the number of words per AS-Unit (**Table 4** and **Figure 5**), the film retelling had the highest value (8.28), followed by the interview (7.12). The other three tasks elicited rather shorter units whose mean length ranged from 4.51 to 5.74 words.

Table 3: Morphological Complexity Index (MCI-10) across tasks.

Task	Mean	SD
Calls	12.75	1.59
Film	11.24	1.46
Neg	12.73	0.90
Map	8.54	1.36
Interv	12.64	1.04

Table 4: Words/AS-Unit across tasks.

Tasks	Mean	SD
Calls	5.01	0.70
Film	8.28	1.51
Neg	4.51	0.83
Map	5.74	1.34
Interv	7.12	1.47

Table 5: Dependent clauses/AS-Unit across tasks.

Tasks	Mean	SD
Calls	0.16	0.05
Film	0.51	0.20
Neg	0.14	0.09
Map	0.11	0.05
Interv	0.31	0.13

The dependent clauses per AS-Unit ratio shows a similar picture, with even more marked differences (**Table 5** and **Figure 6**): While in the film retelling about half of the AS-Units contained dependent clauses, and these were one out of three in the interview, the proportion drops to about 1/8 in the other tasks. Interestingly, the map task implies relatively long syntactic structures, but very little
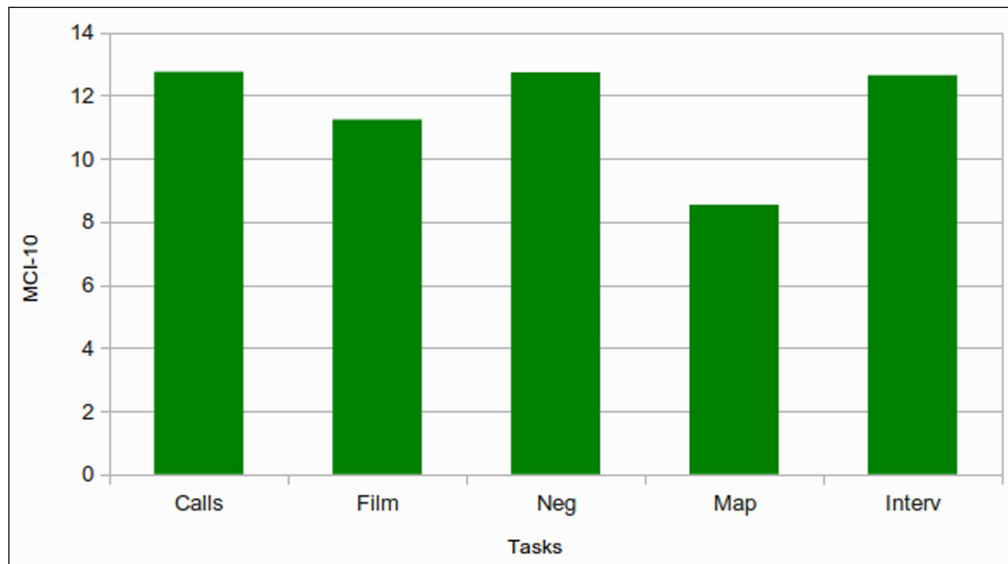
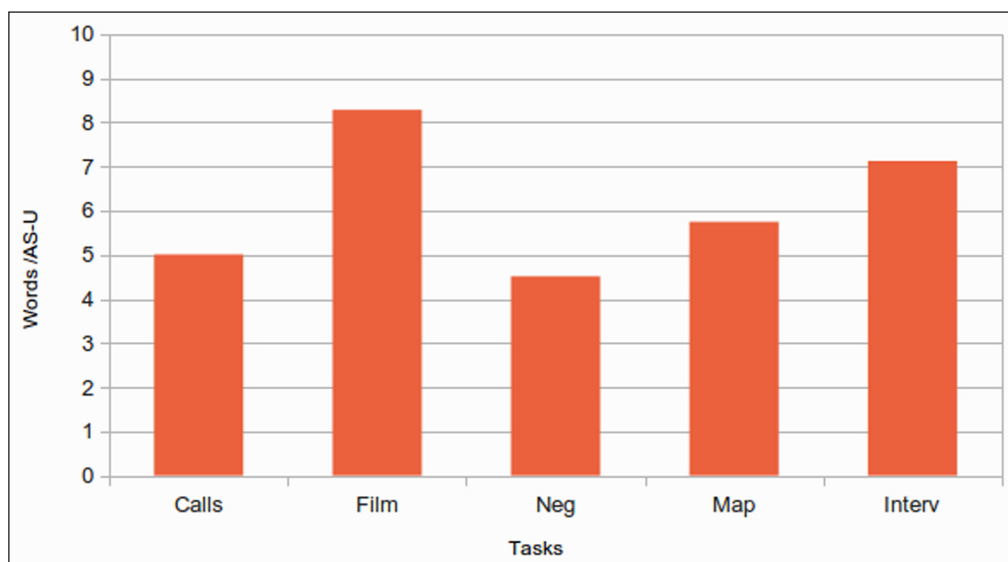**Figure 4:** Morphological Complexity Index (MCI-10) across tasks.
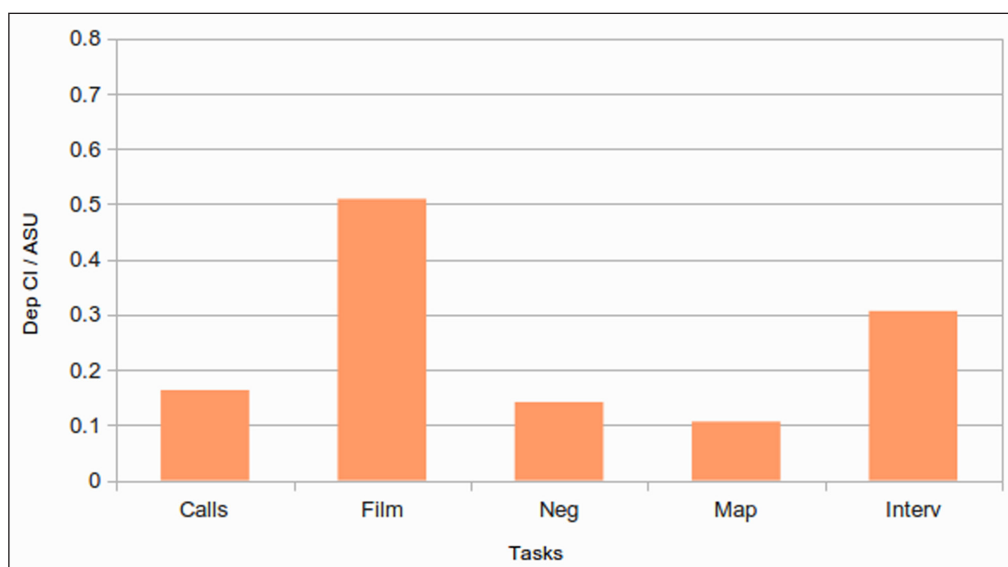

Figure 5: Words/AS-Unit across tasks.

Figure 6: Dependent clauses/AS-Unit across tasks.

subordination: it seems that what is needed to perform it is to construct rich and detailed clauses describing the path and the landmarks, although it does not seem to be necessary to embed other clauses inside them.

4.4. Individual variation

The previous sections reported mean scores achieved by the ten participants across the five tasks. However, it is also important to look at inter-individual variation around these means, to assess whether it differed across tasks. Given that the measures came from different scales, with different value ranges, the coefficient of variation (CV: standard deviation/mean) was used to standardize fluctuations around the mean in order to make them comparable.

What appears from **Table 6** and **Figure 7** is that CV values for lexical diversity and sophistication are rather low across participants and tasks, which means that all participants tended to behave similarly with regards to these dimensions. Variation in the use of morphological processes is slightly higher, especially in some tasks, like the map task, phone calls or film retelling, but still relatively modest. What appears to be highly variable across individuals are syntactic phenomena, which display, in all tasks, high and very high coefficients of variation for both mean length of AS-Unit and the number of dependent clauses per AS-Unit. Syntactic

complexity thus seems to be more related to individual style, allowing for a rather wide range of inter-individual variation, while the lexicon and morphology are more related to task features and less subject to individual preferences.

4.5. Ranking tasks along different dimensions

A final question may be whether there are tasks with high or low levels of all or most dimensions of linguistic difficulty, so that they could be said to be more or less difficult in general, or whether different dimensions vary in a relatively independent manner, so that a task may score high in one and low in another, with a number of possible combinations. To answer this question, **Table 7** sorts tasks in ascending order of difficulty according to the different dimensions considered.

Overall, the map task seems to be relatively easy on most dimensions: It does not require varied lexicon or morphology and it contains the lowest proportion of subordinate clauses. However, compared to other tasks, it elicited a relatively high proportion of infrequent words and its AS-Units were not among the shortest. The interview seemed to imply medium-high use of linguistically difficult structures on all dimensions, viz. lexicon, morphology and syntax, although in no cases did they reach the highest values. For other tasks, the picture is more varied. For instance, retelling the

Table 6: Coefficient of variation across measures and tasks.

Tasks	MATTR	non-basic words	MCI	Words/ AS-U	DepCl/ AS-U
Calls	0.12	0.02	0.12	0.14	0.31
Film	0.07	0.05	0.13	0.18	0.40
Neg	0.08	0.05	0.07	0.18	0.63
Map	0.06	0.12	0.16	0.23	0.48
Interv	0.06	0.07	0.08	0.21	0.41

Table 7: Task difficulty order along different dimension.

MATTR	>2k W	MCI	W/AS-U	DepCl/ AS-U
Map	Negotiation	Map	Negotiation	Map
Calls	Film	Film	Calls	Negotiation
Film	Interview	Interview	Map	Calls
Interview	Map	Negotiation	Interview	Interview
Negotiation	Calls	Calls	Film	Film
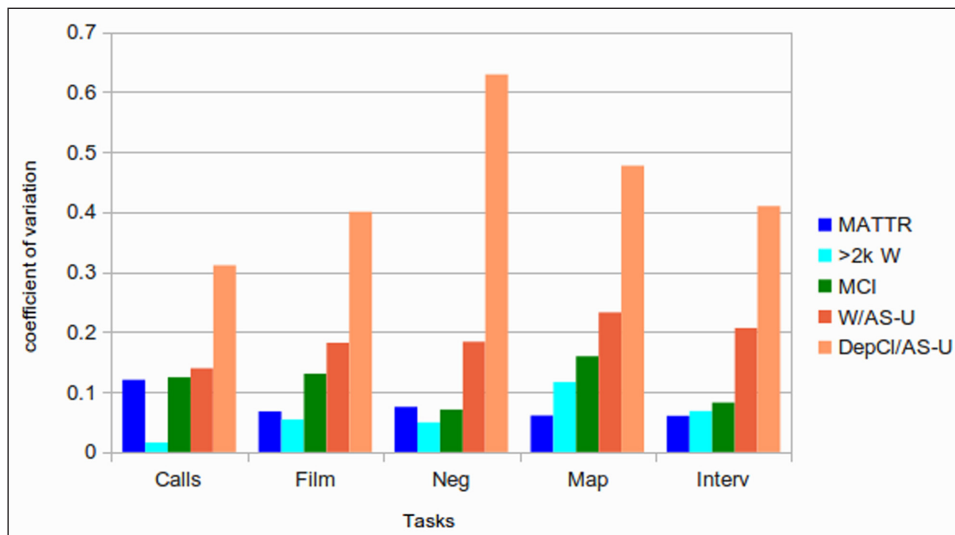
Figure 7: Coefficient of variation across measures and tasks.

video clip implied on the one hand the highest levels of syntactic complexity, with long AS-Units containing a number of dependent clauses; on the other hand, the task could be performed with rather basic vocabulary and little lexical and morphological variation. Phone calls and negotiations offer quite an interesting picture. While both tasks required relatively simple syntactic constructions but a high degree of morphological variety, they sharply differed as regards the lexicon. Phone calls elicited a high number of infrequent words, but the lexicon was rather monotonous, as evidenced by the low MATTR-250 value. The opposite occurred with negotiations, where words were very varied (highest MATTR-250 score), but also very frequent (lowest proportion of uncommon words). This provides additional evidence to the claim that lexical diversity and lexical sophistication are indeed separate dimensions (Skehan, 2009). More generally, the various dimensions of potential linguistic difficulty investigated in this study seem to be rather independent of one another, so that a given task may require high levels in one but relatively low levels in another.

5. Discussion and conclusions

It has long been argued that task difficulty is a multidimensional construct, with many factors contributing to it (Brindley, 1987; Candlin, 1987; Ellis, 2018; Nunan, 1989; Robinson, 2001, 2015; Skehan, 1998, 2015). This article has proposed a procedure to empirically assess one of these dimensions, linguistic difficulty. Results show that this construct is in turn multidimensional and that its sub-components – lexicon, morphology, syntax – vary independently of one another, and there may even be variation in the same sub-domain, such as the lexicon, as evidenced by the different profiles of lexical diversity and sophistication. This implies that future research on task features and demands should take this multidimensionality as a starting point, by carefully manipulating difficulty dimensions one by one rather than pursuing a dichotomic view of tasks as being +/– difficult.

Explicit, analytic and empirically grounded definitions of task linguistic difficulty are desirable for several reasons. First of all they are necessary to continue, in a more principled way, research on the interactions between task difficulty and linguistic performance. Secondly, this line of investigation may contribute to Task-Based Language Teaching (TBLT), by offering more solid grounds to determine the linguistic and communicative demands of different tasks, which is a key aspect for syllabus progression (Baralt, Gilabert & Robinson 2014). Indeed, it has been shown that teachers consider linguistic features one of the most relevant aspects in their evaluation of task difficulty (Révész & Gurzynski-Weiss, 2016). Finally, Task-Based Language Assessment (TBLA) is also concerned with establishing whether different tasks have comparable levels of potential linguistic difficulty, in order to ensure uniformity across multiple editions of the same test or to develop appropriate tasks for different proficiency levels (Elder et al., 2002).

The empirical study reported here was a pilot investigation with the main purpose of presenting an empirical approach to the assessment of task difficulty, and its results seem to be encouraging. It was possible to apply the proposed measures and procedures to the data, and analysis confirms the intuition that the selected tasks should have been rather uniform with respect to their linguistic difficulty. The main purpose of the VIP project on task-based language production was to elicit variation along interactional and sociolinguistic dimensions, while keeping linguistic aspects constant (Pallotti et al., 2011). Therefore, the relatively small range of variation found here for several linguistic dimensions should not be interpreted as a limitation of the approach, nor of the tasks used, but, on the contrary, as a validation of their choice. In other words, a procedure to empirically assess task difficulty can be employed not only to prove that tasks are different, but also that they are similar, or equivalent, at least in some respects.

Despite this overall similarity, the analysis shows that, among the tasks investigated, the map task is the easiest for most of the linguistic dimensions examined. Other tasks present a more complex picture, which suggests that different sub-dimensions of linguistic difficulty are relatively independent of one other. Telling a film (at least the film used in this project), for example, seems to require rather broad and complex syntactic constructions, but fairly basic vocabulary and morphology. On the contrary, asking for information over the phone (again, as regards the phone calls used in this study) implies rather telegraphic syntax and repetitive vocabulary, but involves a rich range of morphological exponents and several low-frequency words. It is also worth noting that the map task, which was found here to be the one with the lowest levels of linguistic difficulty, proved to be the task with the highest interactional difficulty in Pallotti's (2019) study. All this confirms the idea that different dimensions of difficulty, linguistic or of other sorts, are relatively independent and can be manipulated autonomously. It is also possible – although it is just a hypothesis awaiting empirical verification – that trade-off effects may occur, so that as one dimension of difficulty increases, others tend to decrease, even in top language performers.

Finally, it is worth reflecting on inter-individual variability. Even in a relatively homogeneous population such as native speakers, not all individuals behave in the same way, as is to be expected (Andringa, 2014; Dąbrowska, 2019). The present study shows that this individual variability, which may be deemed stylistic, seems to be greater in syntax, as some participants tend to prefer broad and complex structures while others typically produce rather short and simple constructions. Variability is much more limited in the areas of vocabulary and morphology, that seem to be more directly linked to the nature of the task and less to individual preferences.

This study has limitations, too, and calls for further research. Observing top language performers, for instance, has the advantage of reducing the influence of a factor such as linguistic competence on task performance. However, the question remains to what extent learners are actually conditioned by a task's linguistic demands, so that their behaviour follows that of top language

performers. This is the well-known distinction between task-as-workplan and task-in-process (Breen 1987): A task may require, by its very nature, the use of a varied and sophisticated vocabulary, or a wide range of morphological processes, but in its concrete realization learners may resort to much simpler forms. In some cases, these more basic alternatives may still allow learners to achieve the task's goals, perhaps with some more effort and in a less efficient way. There might be other cases, however, in which these requirements are essential, so that the lack of linguistic or communicative skills may result in the impossibility of adequately performing the task. It would therefore be necessary to demonstrate the relationship between the use of certain linguistic behaviours and task success, taking into consideration functional adequacy among the criteria for assessing task performance (Kuiken & Vedder, 2018; Pallotti, 2009; Révész, Ekiert et al., 2016). Furthermore, future research should look at how learners' performance more or less closely matches that of top language performers, and how L2 proficiency may systematically mediate this relationship.

Another outstanding issue is whether different dimensions of linguistic difficulty can be added together, in order to obtain a unitary index of linguistic difficulty, as Pallotti (2019) did for interactional difficulty. This would have clear advantages on a practical level but would require a careful examination of the construct validity of a highly multidimensional notion such as "linguistic difficulty".

Finally, the study presented here needs to be replicated on larger samples and different languages, observing a greater number of potentially difficult linguistic dimensions and employing several tasks. In particular, differences between oral and written productions should be explored and, for each modality, tasks should be controlled for register and genre. This may lead to the inclusion of other measures, for instance phrase length, that have been claimed to be more relevant for assessing syntactic variation in contexts such as academic writing (Biber, Gray & Poonpon, 2011; Ortega, 2012).

Despite these limitations, the present study can be seen as a first attempt at developing a principled, empirically-based procedure to establish which tasks imply higher or lower levels of linguistic difficulty, going beyond current models that assume, on a theoretical level, that this dimension has an impact on global task difficulty, but do not indicate specific methodologies to quantitatively assess it. It also contributes to the debate on how task difficulty may be operationalized and measured, complementing and extending current endeavours based on different methodologies (e.g. subjective perceptions, raters' intuitions, dual task performance, as proposed by Révész, 2014; Révész, Michel et al., 2016; Sasayama, 2016); all these approaches, taken together, will provide a fuller picture of task demands and their potential effects on language performance by native and non-native speakers.

Note

[1] Robinson (2001, 2011) calls difficulty only the challenges that a task poses to a specific individual,

while he uses the term complexity to refer to both a task's structural features (e.g. number of elements) and the cognitive processes it requires of everyone (e.g. spatial or causal reasoning). This terminology however is confusing, as it employs different terms, *difficulty* and *complexity*, for similar constructs (challenges for a given person vs for everyone) and the same term, *complexity*, to refer to different constructs, such as a task's structural properties and the cognitive demands it makes on performers (Skehan, 2015).

References

Abrahamsson, N., & **Hyltenstam, K.** (2009). Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language Learning, 59*(2), 249–306. DOI: https://doi.org/10.1111/j.1467-9922.2009.00507.x

Andringa, S. (2014). The use of native speaker norms in critical period hypothesis research. *Studies in Second Language Acquisition, 36*(3), 565–596. DOI: https://doi.org/10.1017/S0272263113000600

Baralt, M., Gilabert, R., & **Robinson, P.** (2014). *Task sequencing and instructed second language learning*. London: Bloomsbury.

Biber, D., Gray, B., & **Poonpon, K.** (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly, 45*, 5–35. DOI: https://doi.org/10.5054/tq.2011.244483

Breen, M. (1987). Learner contributions to task design. In D. Murphy & C. Candlin (Eds.), *Language learning tasks* (pp. 23–46). Englewood Cliffs, NJ: Prentice-Hall.

Brezina, V., & **Pallotti, G.** (2015). Morphological complexity tool. Retrieved from http://corpora.lancs.ac.uk/vocab/analyse_morph.php

Brezina, V., & **Pallotti, G.** (2019). Morphological complexity in written L2 texts. *Second Language Research, 35*(1), 99–119. DOI: https://doi.org/10.1177/0267658316643125

Brindley, G. (1987). Factors affecting task difficulty. In David Nunan (Ed.), *Guidelines for the development of curriculum resources* (pp. 45–56). Adelaide: National Curriculum Resource Centre.

Bulté, B., & **Housen, A.** (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency—Investigating Complexity, Accuracy and Fluency in SLA* (pp. 21–46). Amsterdam: Benjamins. DOI: https://doi.org/10.1075/lllt.32.02bul

Candlin, C. N. (1987). Towards task-based language learning. In C. Candlin & D. Murphy (Eds.), *Language Learning Tasks* (pp. 5–22). Lancaster: Lancaster University.

Chini, M. (2003). Le phénomène de la jonction interpropositionnelle dans la narration en italien l2: Entre agrégation et intégration. *Acquisition et interaction en langue étrangère, 19*, 71–106.

Collins, L., Troimovich, P., White, J., Cardoso, W., & **Horst, M.** (2009). Some input on the easy/difficult

grammar question: An empirical study. *The Modern Language Journal, 93*(3), 336–353. DOI: https://doi.org/10.1111/j.1540-4781.2009.00894.x

Covington, M., & **McFall, J.** (2010). Cutting the Gordian knot: The Moving-Average Type–Token Ratio (MATTR). *Journal of Quantitative Linguistics, 17*(2), 94–100. DOI: https://doi.org/10.1080/09296171003643098

Dąbrowska, E. (2019). Experience, aptitude, and individual differences in linguistic attainment: A comparison of native and nonnative speakers. *Language Learning, 69*(S1), 72–100. DOI: https://doi.org/10.1111/lang.12323

De Clercq, B. (2015). The development of lexical complexity in second language acquisition: A cross-linguistic study of L2 French and English. *Eurosla Yearbook, 15*(1), 69–94. DOI: https://doi.org/10.1075/eurosla.15.03dec

De Clercq, B., & **Housen, A.** (2017). A cross-linguistic perspective on syntactic complexity in L2 development: Syntactic elaboration and diversity. *Modern Language Journal, 101*, 315–334. DOI: https://doi.org/10.1111/modl.12396

De Clercq, B., & **Housen, A.** (2019). The development of morphological complexity: A cross-linguistic study of L2 French and English. *Second Language Research, 35*(1), 71–97. DOI: https://doi.org/10.1177/0267658316674506

DeKeyser, R. M. (2005). What makes learning second-language grammar difficult? A review of issues. *Language Learning, 55*(S1), 1–25. DOI: https://doi.org/10.1111/j.0023-8333.2005.00294.x

De Mauro, T. (2016). Il Nuovo vocabolario di base della lingua italiana. Retrieved from Internazionale website: https://www.internazionale.it/opinione/tullio-de-mauro/2016/12/23/il-nuovo-vocabolario-di-base-della-lingua-italiana

Dóczi, B., & **Kormos, J.** (2016). *Longitudinal developments in vocabulary knowledge and lexical organization.* DOI: https://doi.org/10.1093/acprof:oso/9780190210274.001.0001

Elder, C., Iwashita, N., & **McNamara, T.** (2002). Estimating the difficulty of oral proficiency tasks: what does the test-taker have to offer? *Language Testing, 19*(4), 347–368. DOI: https://doi.org/10.1191/0265532202lt235oa

Ellis, D. (2011). *The role of task complexity in the linguistic complexity of native speaker output* (Qualifying paper, PhD in Second Language Acquisition Program). University of Maryland.

Ellis, R. (2018). *Reflections on task-based language teaching.* Bristol: Multilingual Matters. DOI: https://doi.org/10.21832/9781788920148

Foster, P., & **Tavakoli, P.** (2009). Native speakers and task performance: Comparing effects on complexity, fluency, and lexical diversity. *Language Learning, 59*(4), 866–896. DOI: https://doi.org/10.1111/j.1467-9922.2009.00528.x

Foster, P, Tonkyn, A., & **Wigglesworth, G.** (2000). Measuring spoken language: a unit for all reasons. *Applied Linguistics, 21*(3), 354–375. DOI: https://doi.org/10.1093/applin/21.3.354

Fulcher, G., & **Márquez Reiter, R.** (2003). Task difficulty in speaking tests. *Language Testing, 20*(3), 321–344. DOI: https://doi.org/10.1191/0265532203lt259oa

Granena, G., & **Long, M. H.** (2013). Age of onset, length of residence, language aptitude, and ultimate L2 attainment in three linguistic domains. *Second Language Research, 29*(3), 311–343. DOI: https://doi.org/10.1177/0267658312461497

Housen, A. (in press). Difficulty and complexity of language features and second language instruction. In C. A. Chapelle (Ed.), *The concise encyclopedia of applied linguistics.* New York: Wiley.

Housen, A., & **Simoens, H.** (2016). Introduction: Cognitive perspectives on difficulty and complexity in L2 acquisition. *Studies in Second Language Acquisition, 38*(2), 163–175. DOI: https://doi.org/10.1017/S0272263116000176

Hulstijn, J. (2019). An individual-differences framework for comparing nonnative with native speakers: Perspectives from BLC Theory. *Language Learning, 69*(1), 157–183. DOI: https://doi.org/10.1111/lang.12317

Hulstijn, J. H. (2015). *Language proficiency in native and non-native speakers.* Amsterdam: Benjamins. DOI: https://doi.org/10.1075/lllt.41

Kang, O. (2013). Linguistic analysis of speaking features distinguishing general English exams at CEFR levels. *Cambridge English: Research Notes, 52*, 40–48.

Kuiken, F., & **Vedder, I.** (2018). Assessing functional adequacy of L2 performance in a task-based approach. In N. Taguchi & Y. Kim (Eds.), *Task-based approaches to teaching and assessing pragmatics* (pp. 266–285). Amsterdam: Benjamins. DOI: https://doi.org/10.1075/tblt.10.11kui

Kuiken, F., Vedder, I., Housen, A., & **De Clercq, B.** (2019). Variation in syntactic complexity: Introduction. *International Journal of Applied Linguistics,* (1– 10). Retrieved from DOI: https://doi.org/10.1111/ijal.12255

Laufer, B. (1995). Beyond 2000. A measure of productive lexicon in a second language. In L. Eubank, L. Selinker & M. Sharwood Smith (Eds.), *The current state of interlanguage* (pp. 265–272). Amsterdam: Benjamins. DOI: https://doi.org/10.1075/z.73.21lau

Long, M. (2015). *Second language acquisition and task-based language teaching.* Malden, MA: Wiley-Blackwell.

Malicka, A., & **Levkina, M.** (2012). Measuring task complexity: Does L2 proficiency matter? In C. Coombe & A. Shehadeh (Eds.), *Task-based language teaching in foreign language contexts: Research and implementation* (pp. 43–66). Amsterdam: Benjamins. DOI: https://doi.org/10.1075/tblt.4.06mal

Norris, J. M., & **Ortega, L.** (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*(4), 555–578. DOI: https://doi.org/10.1093/applin/amp044

Nunan, D. (1989). *Designing tasks for the communicative classroom.* Cambridge: Cambridge University Press.

Ortega, L. (2012). Interlanguage complexity: A construct in search of theoretical renewal. In B. Kortmann & B. Szmrecsanyi (Eds.), *Linguistic complexity: Second*

language acquisition, indigenization, contact (Vol. 13, pp. 127–155). Berlin: De Gruyter.

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, *30*(4), 590–601. DOI: https://doi.org/10.1093/applin/amp045

Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research*, *31*(1), 117–134. DOI: https://doi.org/10.1177/0267658314536435

Pallotti, G. (2019). Assessing tasks: The case of interactional difficulty. *Applied Linguistics*, *40*(1), 176–197. DOI: https://doi.org/10.1093/applin/amx020

Pallotti, G., Ferrari, S., & Nuzzo, E. (2011). A systematic procedure for assessing communicative competence. In G. Videsott & W. Wiater (Eds.), *New theoretical perspectives in multilingualism research* (pp. 113–133). Bern: Peter Lang.

Plonsky, L., & Kim, Y. (2016). Task-based learner production: A substantive and methodological review. *Annual Review of Applied Linguistics*, *36*, 73–97. DOI: https://doi.org/10.1017/S0267190516000015

Rescher, N. (1998). *Complexity: A philosophical overview.* New Brunswick, NJ: Transaction Publishers.

Révész, A. (2014). Towards a fuller assessment of cognitive models of task-based learning: Investigating task-generated cognitive demands and processes. *Applied Linguistics*, *35*(1), 87–92. DOI: https://doi.org/10.1093/applin/amt039

Révész, A., Ekiert, M., & Torgersen, E. N. (2016). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics*, *37*(6), 828–848. DOI: https://doi.org/10.1093/applin/amu069

Révész, A., & Gurzynski-Weiss, L. (2016). Teachers' perspectives on second language task difficulty: Insights from think-alouds and eye tracking. *Annual Review of Applied Linguistics*, *36*, 182–204. DOI: https://doi.org/10.1017/S0267190515000124

Révész, A., Michel, M., & Gilabert, R. (2016). Measuring cognitive task demands using dual-task methodology, subjective self-ratings, and expert judgments: A validation study. *Studies in Second Language Acquisition*, *38*(4), 703–737. DOI: https://doi.org/10.1017/S0272263115000339

Robinson, P. (1995). Task complexity and second language narrative discourse. *Language Learning*, *45*(1), 99–140. DOI: https://doi.org/10.1111/j.1467-1770.1995.tb00964.x

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics*, *22*, 27–57. DOI: https://doi.org/10.1093/applin/22.1.27

Robinson, P. (2011). Second language task complexity, the Cognition Hypothesis, language learning, and performance. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance* (pp. 3–38). Amsterdam: Benjamins. DOI: https://doi.org/10.1075/tblt.2

Robinson, P. (2015). The Cognition Hypothesis, second language task demands, and the SSARC model of pedagogic task sequencing. In M. Bygate (Ed.), *Domains and directions in the development of TBLT* (pp. 87–121). Amsterdam: Benjamins. DOI: https://doi.org/10.1075/tblt.8.04rob

Sasayama, S. (2016). Is a 'complex' task really complex? Validating the assumption of cognitive task complexity. *The Modern Language Journal*, *100*(1), 231–254. DOI: https://doi.org/10.1111/modl.12313

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing.* Manchester, UK.

Skehan, P. (1992). Strategies in second language acquisition. *Thames Valley University Working Papers*, *1*.

Skehan, P. (1998). *A cognitive approach to language learning.* Oxford: Oxford University Press. DOI: https://doi.org/10.1177/003368829802900209

Skehan, P. (2009). Lexical performance by native and non-native speakers on language-learning tasks. In B. Richards (Ed.), *Vocabulary studies in first and second language acquisition* (pp. 107–124). London: Palgrave Macmillan. DOI: https://doi.org/10.1057/9780230242258_7

Skehan, P. (2015). Limited attentional capacity and cognition: Two hypotheses regarding second language performance on tasks. In M. Bygate (Ed.), *Domains and directions in the development of TBLT* (pp. 123–155). Amsterdam: Benjamins. DOI: https://doi.org/10.1075/tblt.8.05ske

Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. *Language Learning*, *60*(2), 263–308. DOI: https://doi.org/10.1111/j.1467-9922.2010.00562.x

Treffers-Daller, J. (2013). Measuring lexical diversity among L2 learners of French. In S. Jarvis & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 79–104). Amsterdam: Benjamins. DOI: https://doi.org/10.1075/sibil.47.05ch3

Vercellotti, M. L. (2018). Finding variation: assessing the development of syntactic complexity in ESL Speech. *International Journal of Applied Linguistics*, 1–15. DOI: https://doi.org/10.1111/ijal.12225

Wen, Z., & Ahmadian, M. (Eds.) (2019). *Researching L2 task performance and pedagogy.* Amsterdam: John Benjamins. DOI: https://doi.org/10.1075/tblt.13

Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied Linguistics*, *31*(2), 236–259. DOI: https://doi.org/10.1093/applin/amp024