This is the author's manuscript, after peer review and before copy-editing and typesetting. The article's final version is published here: Gabriele Pallotti, Assessing Tasks: The Case of Interactional Difficulty, *Applied Linguistics*, Volume 40, Issue 1, February 2019, Pages 176–197, https://doi.org/10.1093/applin/amx020

Assessing Tasks: The Case of Interactional Difficulty

Gabriele Pallotti
University of Modena and Reggio Emilia

INTRODUCTION

The notions of task difficulty and task complexity have received considerable attention from researchers in Second Language Acquisition (SLA), Task-Based Language Teaching (TBLT) and Task-Based Language Assessment (TBLA). In SLA research, the main question has been whether different types and levels of task difficulty are related to systematic variations in L2 task performance (for an overview, see Long 2015). A recurring theme in TBLT concerns criteria for grading tasks, in order to assign students activities that are hard enough to promote learning but not so hard as to become frustrating (e.g. Baralt, Gilabert and Robinson 2014; Robinson 2011). This concern has also been shared by proponents of TBLA, who need to know which tasks are most appropriate for assessing learners at different proficiency levels, or whether task difficulty has any systematic correlation with test scores (e.g. Brown, Hudson, Norris and Bonk 2002; Shehadeh 2012).

However, recent critical overviews (Long 2015; Norris 2016; Révész 2014) have pointed out that, despite the high number of studies looking at task 'difficulty' or 'complexity', there is still considerable controversy about the meaning of these constructs, how they can be operationalized and how such operationalizations may be validated, which calls for more theoretical and methodological work.

This article intends to contribute to this debate on three different levels. Terminologically, it will argue that what many SLA studies thus far have called 'task complexity' should be more appropriately called 'task difficulty'. Theoretically, it will present a new construct, *task interactional difficulty*, and will discuss ways of operationalizing it that rely on socio-interactionist notions such as turn-taking and participation dynamics, rather than on the cognitive-interactionists constructs that have been prevailing in previous SLA research on tasks. Methodologically, it will discuss a new way of assessing task demands based on native speakers' behavior.

The article's main contribution are, thus, conceptual and methodological as it discusses some key constructs used in task-based research and suggests some more refined practical operationalizations. These proposals are applied to a relatively small sample of participants and tasks, so that the empirical research reported here should be seen as a demonstration of the feasibility of the approach rather than as a validation study proper.

COMPLEXITY OR DIFFICULTY?

Although a considerable number of studies in applied linguistics and other areas of the social sciences have been concerned with task complexity, 'it is hard to imagine any other construct could equal task complexity in terms of the level of ambiguity and internal inconsistency achieved over the years' (Gill and Hicks 2006: 2). Its multiple meanings may be grouped in three broad 'perspectives' (Campbell 1988):

- the psychological experience perspective;

- the task-person interaction perspective;
- the objective characteristics perspective.

In his framework for investigating tasks in SLA research, Robinson (2011) proposes to use the term ’complexity’ for the first perspective, that is, the features of a task making it more cognitively demanding for performers in general, regardless of their individual characteristics. He calls ’difficulty’ the second perspective, resulting from the encounter of certain task features with an individual’s personal endowment (e.g., previous knowledge, skills, aptitude, working memory). The third perspective, which has to do with the internal structure of a task, has not been systematically kept apart in SLA research, where it has usually been considered as contributing to cognitive demands. Thus, for example, ’number of elements’, a structural task feature, appears in Robinson’s taxonomy together with ’spatial, causal and intentional reasoning’ or ’planning time’, which clearly have to do with participants’ cognitive processes.

Skehan (1998: 134), on the other hand, calls ’difficulty ... the level of challenge that a task is likely to contain’, which corresponds to Robinson’s ’complexity’. The same usage is found in Tavakoli (2009) and in the language testing literature, where ’difficulty’ is always used instead of ’complexity’ (e.g. Fulcher and Márquez Reiter 2003; Brown, Hudson, Norris, Bonk 2002; Taguchi 2007). In another publication, Skehan and Foster (2001) seem to employ ’complexity’ and ’difficulty’ virtually as synonyms, like R. Ellis (2003: 351), who defines task complexity as ’the extent to which a particular task is inherently easy or difficult’.

This terminological uncertainty is not an ideal starting point for developing a construct’s theoretical and operational definition. In this article we will follow Pallotti (2009; see also Housen and Simoens 2016 for similar choices in their discussion of linguistic difficulty and complexity), who suggests to restrict use of the term *complexity* to the third perspective, that is, to a task’s inherent, structural, characteristics, like the number of elements and their relationships, and to call *difficulty* the first and second perspectives. These can in turn be differentiated as ’interindividual’ (or task-inherent) and ’individual’ (or person-inherent) difficulty: playing Lizst’s *Transcendental études* is interindividually more difficult, i.e. more difficult for everyone than playing *Twinkle, Twinkle, Little Star*, although a beginner pianist will find both more difficult than an acclaimed performer. More particularly, here we will be concerned with the first perspective only, ’interindividual difficulty’, that is the difficulty inherent in the task itself, and not with the difficulty encountered by specific individuals with particular endowments. The ’interindividual difficulty’ which is the focus of this article thus corresponds to what Robinson and others would call ’complexity’. The term ’difficulty’ is in our opinion more transparent and less ambiguous, inasmuch as it better represents the similarities between two closely related notions, such as individual (for someone) *vs* interindividual (for everyone) difficulty, rather than calling them by two completely different names - ’complexity’ and ’difficulty’ - as in Robinson’s framework. The term ’complexity’ can thus be reserved to describe a task’s objective characteristics, and discovering whether different amounts of structural complexity pose more cognitive demands to all or some individuals becomes an empirical question in its own right.

PREVIOUS RESEARCH

Defining and operationalizing task difficulty

A number of SLA studies have been conducted on the relationships between task difficulty (often deemed ’complexity’) and L2 performance, described in most cases in terms of linguistic Complexity, Accuracy and Fluency (CAF), or of interactional modifications, like negotiation of meaning and language-related episodes. After an extensive review of the field, Long (2015: 244) concludes that ’clear and consistent findings are few and far between’. Among the reasons he gives, the most prominent are methodological – both task characteristics (independent variables) and

performance on task (dependent variables like CAF) have been characterized in too many different ways, often with little concern about construct definition and operationalization, making it impossible to cumulatively compare studies and their results.

Most early studies on task difficulty in SLA took what might be called an *a priori* approach. A 'simple' and a 'complex' version of the same task were compared (85% of the studies reviewed by Sasayama, Malicka and Norris 2015, cited in Sasayama 2016, formulated their research question in these dichotomous terms), with complexity being operationalized along one or several dimensions, such as those listed by Robinson (2011), e.g. ± few elements, ± reasoning demands, ± planning time. While it may be relatively straightforward to determine, in a single study, what is more or less complex, it is not clear how these dichotomies are to be operationalized in general, absolute terms. For instance, how many elements are 'few'? And what is an element, exactly (Kuiken and Vedder 2007; D. Ellis 2011)? One could even conceive of a study where the simple condition involves 3 elements and the complex one 6, and of another where simple means 6 elements and complex 12, so that 6 elements would end up being complex in one case and simple in another.

A second methodological issue is that most of this research assumed, rather than demonstrated, that changes along these task dimensions do indeed bring about higher cognitive demands, that is, more difficulty. However, if increasing the number of elements in one version of a task does not produce any difference in performance (e.g. in lexical diversity), it is not clear whether this should be interpreted as evidence against the model positing a relationship between task difficulty and linguistic performance, or rather as showing that adding more elements does not increase task difficulty.

To overcome this problem, some studies have sought independent evidence of task difficulty. One of the most common and practical ways to do so has been collecting participants' self-ratings, both through interviews and questionnaires directly asking which tasks were felt to be more difficult, or with indirect measures, such as the perception of time spent on task (see Baralt, Gilabert and Robinson 2014 for a review). A few more recent studies have added more sophisticated measures, triangulating several of them at a time, to validate the claim that a task is more difficult than another. For example, Révész, Sachs and Hama (2014) collected experts' ratings and looked at the number and length of eye fixations and performance on a secondary task (dual-task methodology) as further independent measures of cognitive load. Révész, Michel and Gilabert (2016) validated the difference between ± complex versions of three tasks by analyzing participants' self-ratings, their performance on a secondary task and experts' judgments. Sasayama (2016) employed dual-task methodology, subjective time estimation, and self-ratings to assess the difficulty of four narrative tasks involving an increasing number of elements. She found that only the most extreme versions (the simplest and the hardest) produced reliable differences, and that the effects of task difficulty varied across L2 proficiency levels.

This last finding by Sasayama raises a third methodological question. Most research about task effects on linguistic performance has been based on learner production and perception data; while this choice may seem natural, it also creates further problems. Let us suppose again a study where an increase in task difficulty does not produce any variation in learners' lexical diversity. Besides the two options considered above – (a) there is no relationship between task difficulty and lexical diversity; (b) the task was not more difficult – there would be a third possible explanation: (c) the impact of task difficulty on lexical diversity does not manifest itself *in this particular type of learner*. Besides Sasayama's research, other studies have shown that task difficulty effects on performance do vary across learners at different proficiency levels, for example as regards CAF measures (Malicka and Levkina 2012) or interactional dynamics (Gilabert, Barón and Llanes 2009; Kim 2009).

To overcome these problems, another way of assessing task difficulty independently of learners' proficiency levels may be looking at native speakers. As Long (2015: 239) notes, observing native speakers' performance 'in task complexity research, initially, at least, offers a

simple way of controlling for accuracy, processing demands and (to a lesser extent) fluency, thereby allowing any changes in the one remaining dependent variable, linguistic complexity, to be isolated as the effect of changes in task complexity. Why search in the dark for that relationship in non-natives before first ascertaining its existence in natives?'

Among the SLA studies that have looked at native speakers' task performance, Foster and Tavakoli (2009) and D. Ellis (2011) show that different tasks or versions of the same task systematically lead to different levels of syntactic complexity, while Michel, Gilabert and Révész (2014) and Révész, Michel and Gilabert (2016) observed that native speakers' eye movements and their accuracy in a secondary task were systematically correlated to the main task's complexity. As regards the ± interaction variable, Michel (2011) reports that while L2 learners production showed increased lexical diversity in the interactive condition, exactly the reverse occurred for native speakers. This was probably due to the non-natives being able to copy each other's words in dialogues, thus enriching their lexicon, while natives did not need help of this kind showing once again how complicated it is to tease apart task effects from speakers' (in)competences.

Task difficulty and interaction

A number of studies have looked at how task difficulty impacts on interactional features like clarification requests, confirmation checks or recasts (e.g. Gilabert et al. 2009; Kim 2009; Révész 2011). Much less research has taken interaction as an independent variable, that is, as a task feature bearing on linguistic performance. Among these studies, Michel et al. (2007) and Michel (2011) have compared tasks that were similar in all respects except for the ± monologic condition. They report that, in dialogues, both native and non-native speakers tend to produce more accurate and fluent, but syntactically less complex, utterances. In Michel (2011), non-native speakers' dialogues were more fluent, accurate and lexically varied than monologues, although for native speakers lexical variety was higher in monologues; Michel et al. (2007), however, did not find any effect of interaction on this dependent variable. Gilabert, Barón and Levkina (2011) also found that learners tended to produce more fluent but syntactically less complex utterances in interactive tasks; proficiency differences among learners had a clear impact on monologic performance, which disappeared in dialogues, probably because participants could rely on each other's turns in constructing their utterances.

To summarize, only a few studies thus far have attempted to assess task difficulty independently of L2 task performance, and most of them have been concerned with cognitive aspects, such as perceived effort or the amount of mental processing, as evidenced e.g. by eye movements or performance on a secondary task. Interaction, too, has been typically seen as a dependent variable related to cognitive processes like noticing and negotiation of meaning; the very few studies treating it as an independent variable operationalized it in dichotomous terms (absent vs present), with no clear results about its contribution to overall task difficulty.

This article intends to contribute to the current debate by addressing the following issues:

- How can task interactional difficulty be theoretically and operationally defined in terms of socio-interactional features like turn-taking and participation dynamics?
- Can the construct be practically applied to a sample of native speakers performing several tasks?
- Is it possible to develop an index which can be used to rank tasks on an ordinal scale of interactional difficulty, thus going beyond dichotomies like ± interaction or ± difficult?
- Can this procedure be extended to other facets of task difficulty?

TASK INTERACTIONAL DIFFICULTY: THEORETICAL DEFINITION

None of the studies looking at interaction as a task feature explicitly addressed the issue of

'task interactional difficulty', which remains, to the best of our knowledge, an unexplored construct. In order to provide a working definition, the meaning of the three terms needs to be clarified.

A **task** can be defined as 'an activity in which meaning is primary; there is some communication problem to solve; there is some sort of relationship to comparable real-world activities; task completion has some priority; the assessment of the task is in terms of outcome' (Skehan 1998: 95).

Interactional: Interactional competence may be defined as that competence which is *specifically* needed to inter-act, that is to participate in courses of action involving two or more speakers, and which crucially has to do with turn-taking dynamics. This definition is thus narrower than those usually proposed by scholars working in the socio-interactionist approach, who tend to see interactional competence as comprising all the competences needed for verbal interaction.

Hall and Pekarek Doehler (2011: 2), for instance, list 'knowledge of social-context-specific communicative events or activity types, ... the ability to deploy and recognize context-specific patterns by which turns are taken, actions are organized and practices are ordered. And ... the prosodic, linguistic, sequential and nonverbal resources conventionally used for producing and interpreting turns and actions, to construct them so that they are recognizable for others, and to repair problems in maintaining shared understanding of the interactional work we and our interlocutors are accomplishing together'. Such wide-ranging definitions, while valid in principle from a theoretical point of view, make it impossible to establish what is really peculiar to interactional competence and how it can be differentiated from and related to other aspects of communicative competence. On the other hand, our definition covers more ground than is usually covered in cognitive-interactionist research (for a review, see Mackey 2012), where 'interaction' is typically operationalized in terms of the conversational sequences used to address communicative problems, like clarification requests, confirmation checks, recasts or other language-related episodes. However, it is clear that, from a theoretical point of view, interaction involves more than interacting to solve linguistic and communicative problems.

Difficulty. Given the extant uncertainty in the scientific literature regarding terms like *difficulty* or *complexity*, it is perhaps advisable to start from some relatively neutral dictionary definitions. *Difficult* can be defined as 'needing skill or effort' (Cambridge Dictionary), 'requiring much work or skill to do or make' (Merriam Webster's), or 'needing much effort or skill to accomplish, deal with, or understand' (Oxford Dictionary). All these definitions mention *effort* (or *work*) and *skill*. The two basic dimensions of effort and skill also appear in Housen and Simoens' definition of linguistic difficulty: 'a language feature is more difficult than another if its processing and learning requires *more time* and/or *more mental activity* from a particular language learner in a particular learning context' (2016: 166; emphasis added).

Empirically, *effort* can be gauged with introspective reports, subjective perceptions of time spent on task, dual task methodology and physiological measures such as heart, brain or eye activity (e.g. Révész 2014; Révész et al 2015; Sasayama 2016).[1] The amount of *skill* needed to successfully perform a task or action can be assessed either by asking experts, such as language teachers, or by looking at the time it takes an average person to reach a satisfactory performance level, as documented by developmental studies on the acquisition of various aspects of language ability.

[1] Révész et al (2015) found that participants' ratings for their perceptions of a task being 'difficult' or 'requiring mental effort' were highly correlated, and the correlation was even stronger for expert judges.

TASK INTERACTIONAL DIFFICULTY: OPERATIONAL DEFINITION

Defining task difficulty based on native speakers' performance

Having defined interactional difficulty from a theoretical point of view, its operational definition can be based on the following 'validity argument' (Kane 2013), which may apply to other sources of difficulty as well.

Firstly, some behaviors are deemed to be more difficult, in terms of skill, because they are mastered late in the course of acquisition. For instance, research on lexical development has shown that increased lexical diversity or a high proportion of rare words characterize advanced stages of L2 acquisition (Milton 2009). Similarly, some grammatical structures are typically acquired late, which may be taken as evidence that they are grammatically more difficult (Housen and Simoens 2016). Some communicative behaviors, too, can be said to be interactionally more difficult, as will be shown in the next section.

Secondly, tasks are assessed to establish to what extent they require these difficult behaviors, or, in other words, their expected level of 'code complexity' (Skehan 1998). To do so, one may interview experts such as teachers or linguists. While this source of evidence is valuable and can be triangulated with others, it has the limitation of being based on *a priori* intuitions, rather than on the observation of real people performing the task.

However, looking at task performance raises some issues, too. In fact, if the task were performed by a person with limited competences, it would be impossible to ascertain whether the absence of difficult linguistic-communicative behaviors would be due to the task not requiring them or to the task taker's limitations. Thus, the ideal task taker in this regard would be one with the highest possible competence level, that is, one mastering a very wide range of linguistic structures and communicative practices, including those known to be difficult for L2 learners. It is both practical and reasonable to identify this ideal task taker with a native speaker. Clearly, native speakers do not form a completely homogeneous population, and their linguistic performance does vary, both because of inevitable individual differences in terms of personality or cognitive style, but also because of attributes having more systematic effects, such as age and socio-educational background (Hulstijn 2015). However, once these last sources of systematic variance are controlled by selecting natives with profiles similar to the target L2 learners, including their knowledge of the relevant content domain (Zuengler and Bent 1991), there is ample evidence showing that - especially for the most common structures forming our 'basic language cognition' (Hulstijn 2015) - native speakers' performance is much more uniform than that of L2 learners, even very advanced learners, and that the former invariably score at the highest level on all measures (Granena and Long 2013). This does not exclude the possibility that some L2 users may reach this level, which is an empirical question in its own right. However, from a practical point of view, selecting a group of native speakers is a relatively straightforward way to ensure that these participants will be at ceiling in their command of the basic linguistic forms and functions involved in most tasks employed in SLA research, TBLT and TBLA. Thus, their variable use across tasks of structures that are more or less difficult for L2 acquirers can be said to be primarily due to the tasks themselves, rather than to the participants' (in)competences.

Measures to assess interactional difficulty

Based on the existing literature on L2 acquisition, three parameters appear to be relevant to define difficult interactional behaviors: number of turn exchanges, number of initiating moves and visual contact with the interlocutor.

Number of turn exchanges. Taking and yielding turns is not easy: it requires a considerable degree of coordination among participants to identify 'transition relevant places' (Sacks *et al*. 1974) and exploit them to swiftly insert one's contribution in the conversation's flow. A dialogue in which

turns are frequently exchanged will thus be interactionally more difficult than one in which speakers can maintain the floor for a relatively long time. Some studies have demonstrated that this ability grows slowly in a second language (for a review, see Pekarek-Doehler and Pochon-Berger 2015). For instance, Pallotti and Ferrari (2008) and Nuzzo and Gauci (2012) show that intermediate-advanced learners of L2 Italian, despite their good linguistic skills, tend to favor long and complex turns in telephone calls' openings, thereby displaying their difficulty in managing the rapidly-paced turn exchanges that native speakers frequently produce in the same conditions. Michel (2011: 168), too, observed that while turns in dialogic tasks tend to be generally shorter than in monologues, the difference is more pronounced in native speakers, thus demonstrating that L2 learners are not equally at ease in conversations where short, telegraphic turns are rapidly exchanged.

For practical purposes, a turn can be defined as a spate of talk produced by a single speaker not interrupted by others, and can be anything between a single word and dozens of sentences.

Number of initiating moves. A move can be considered to be initiating if it represents a 'first pair part', i.e. a conversational action establishing a state of 'conditional relevance' whereby the interlocutor is expected to produce a second pair part (Schegloff and Sacks 1973). Questions are a prototypical case of initiating moves, but the category also includes proposals, invitations, greetings. More generally, an initiating move makes the conversation progress, having a proactive function which makes a reactive move relevant.

Research has repeatedly shown that in the early stages of L2 acquisition learners tend to be rather passive and most of the times they speak only when a more competent interlocutor involves them in the conversation. The development of interactional competence has thus been described as a gradual increase in the ability to take initiative, to play a more active role in moving the conversation forward (Cekaite 2007; Nguyen 2011; Pallotti 2001; Young and Miller 2004).

A turn may contain several moves, for example an answer to a previous question (reaction) and a new question, or proposal etc. (initiation). In such cases, the present operationalization would score an initiating move anyway, as the analytic unit is the move, not the turn. Furthermore initiation appears to be a scalar notion, along a continuum of proactivity / reactivity. A question is clearly an initiating move (even if it may follow another question), and an answer to it is clearly a reaction. Proposals, too, can often be easily classified as initiating moves, followed by partial or total acceptances/refusals; if a proposal is followed by another proposal, both are counted as initiating moves.

Visual access. The third factor contributing to a task's interactional difficulty is eye-contact among participants. Conceptually, this dimension is of a different nature from the previous two, as it is a direct reflection of the task's structure and does not require observation of performance. Multimodal communication, flowing through multiple channels, affords a higher degree of redundancy and allows possible linguistic gaps to be compensated through non-verbal resources. Gaze is systematically used to manage turn-taking dynamics (Rossano 2013), and several studies have shown that non-native speakers understand oral messages better when they are accompanied by gestures and facial expressions (Dahl and Ludvigsen 2014; Sueyoshi and Hardison 2005; Wagner 2010). Gullberg (1998, 2006) reports that gestures can also be strategically used by L2 learners to ensure discourse cohesion, to fill lexical gaps and to manage interactional flow.

Other measures

There are other dimensions which may at first sight seem to bear on interactional difficulty, but which on closer inspection turn out to concern conceptually different, albeit related, dimensions.

One is the range of communicative moves. A task that involves asking, answering, proposing, describing, explaining and negotiating intuitively looks more difficult than one in which one must just describe. Another possibly relevant dimension has to do with politeness requirements. Fulcher and Márquez Reiter (2003) and Taguchi (2007), for instance, have shown that tasks implying higher

demands in terms of more power asymmetry, social distance and degree of imposition are perceived to be more difficult by test-takers and lead to less fluent linguistic production. While these remarks are certainly relevant for a discussion of social factors impacting on task difficulty, they concern a general *pragmatic* competence, which is not specifically interactional, for it can be displayed in monologues as well.

The number of participants involved is a parameter more directly linked to interaction, but it is not clear whether it contributes to more or less difficulty. In fact, on the one hand a greater number of participants requires the skill to manage parallel conversations, to compete for the floor with more people, to heed several contributions at the same time. On the other hand, more participants also mean less need to be active and to move on the conversation, which decreases interactional difficulty (Schegloff 2007). Given these contradictory effects, the parameter will not be included in the construct's operational definition.

THE STUDY

Participants and method

Data come from a larger research project on the acquisition and use of Italian as a first and additional language, focusing on girls in their late teens, both L2 learners (N=14) and native speakers (N=10). This study will look at this last group only (mean age = 18.0).

All participants performed a number of communicative activities, so that their linguistic skills could be assessed in a range of contexts with different interlocutors. The procedure consisted in two sessions on two different days. The first session involved a series of tasks commonly used in SLA research and began with a long semi-structured interview with an adult female interviewer. This was followed by retelling a silent film and a picture story, then by a map task in which the girl ('instruction giver') gave instructions to the adult ('instruction follower'), using maps with the same path but with slightly different landmarks, to increase communicative demands and provide opportunities for meaning negotiation. The second session, a few days later, involved more interactive tasks with two or three participants having more balanced communicative roles. There was another map task, this time with the peer. Then two information-seeking activities followed, one requiring 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, but also to a list of 'experts' who could be asked to provide advice and information. In these tasks, the adult interviewer actively participated in the negotiations about action planning and decision-making, so that there were three participants interacting on the scene. Apart from the initial ice-breaking conversation, all the other tasks were presented in random order in different sessions. Data were audio- and video-recorded. More details about the tasks and the corpus can be found in the online Appendix and in Pallotti, Ferrari and Nuzzo (2011).

In this article we will look at native speakers' data from the interview, film retelling, map task with adult and peer and from phone calls and face-to-face negotiations in the school trip's organization. To the best of our knowledge, this is the first study in applied linguistics collecting and comparing data from six different tasks at the same time. Given that interviews and the school trip organization tasks lasted much longer than other activities, only the first ten minutes of the former and the last ten of the latter will be analyzed here.[2]

Thus, the corpus used for this study consists of 60 communicative episodes (6 tasks * 10

[2] These sub-samples were selected as they were more uniform across tasks and participants: the first part of the interviews contained a rather standard set of ice-breaking questions, while the last part of the school trip organization was carried out in a similar way by all pairs, with recurrent

participants), with about 73,200 words and 10,200 turns. Transcription followed a modified version of the Chat-CA system and included pause length, overlaps, voice quality and non verbal behaviors like laughter, sighs, coughs etc. About 60% of the data were first coded by research assistants, 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. The online Appendix contains an excerpt exemplifying data coding and analysis.

Before presenting the results, a few methodological clarifications are in order. First, while turn exchanges were seen as a property belonging to the whole task itself, and were thus counted the same for all participants, initiating moves were computed only for one participant (the *focal participant*) in all tasks except for the negotiation, where both speakers were seen as focal participants. Secondly, given that the number of words produced by each participant in each task varied, all scores have been standardized as values per 100 words. Finally, the median will be used as the main central tendency measure, since it is more in line with the final score's attribution based on quartiles; accordingly, variation in the data will be expressed by the Inter-Quartile Range (IQR). However, descriptive statistics in the tables will also include the mean and standard deviation, whose general trends do not substantially differ from those of median and IQR.

RESULTS

As regards turn exchange frequency (Table 1, Figure 1), the task with fewer turn exchanges is the film retelling, with a median value of 1.96 per 100 words. This value raises to 9.87 for the interview, characterized by a steady exchange of questions and answers, and even more so in telephone calls (12.21). The map task with an adult (18.69), that with a peer (19.5) and the decision-making negotiation (19.69) present very similar scores, with a median value of about one turn exchange every 5 words.

[TABLE 1 AND FIGURE 1 NEAR HERE]

As regards initiating moves (Table 2, Figure 2), both interview and film retelling have medians close to zero (0.05 and 0.18, respectively), as in these tasks the focal participants' communicative role is clearly that of responding - either to many questions in a row in the interview, or to just one general question 'tell me what happened' in the film. They very rarely take the initiative and this basically occurs to ask some clarification questions or to check the interlocutor's comprehension. Initiatives are more frequent in negotiations to decide what to do on the school trip (2.05 per 100 words), even more so in phone calls (3.34) and with much higher proportions (5.04 and 6.36 per 100 words, respectively) in map tasks with adult and peer. This trend can partly be explained by the adult's role in the task. In negotiations, the adult was an 'official participant' (Goffman 1981) to the interaction, who could make suggestions and proposals just like the younger participants; furthermore, both girls could take the initiative, which made the initiating space left to each of them rather limited (which is a further demonstration that participants' number cannot be taken as a factor automatically increasing interactional difficulty, if only because there is a trade-off between participants' number and the possibility, or necessity, for each of them to take the initiative). In phone calls, participants often took an initiating role when asking questions, but these were then followed by rather long answers by their interlocutors, which explains the relatively low proportion of initiating moves per 100 words. Map tasks seem to be the tasks that most favors conversational initiatives - the instruction-giver has the explicit role of directing her partner, and this occurs more frequently with the peer than with the adult, who often took an initiating role by

activities like agreeing on a final decision, taking note of it, and asking the adult partner for confirmation.

asking what she had to do, rather than letting the younger interlocutor explain on her own initiative. This corroborates Brooks' (2009) finding that tasks with two candidates produce more interaction and more varied moves than those with an examiner, whose 'up' role tends to lead to more controlled conversational formats.

[TABLE 2 AND FIGURE 2 NEAR HERE]

As is apparent from the tables and graphs, there is a fair amount of variation in the data. This is unsurprising, given the freedom that was allowed to participants and the well-known variability in people's conversational styles. Nonetheless, this variation was not completely random, with individual scores rarely being very far off the median value, and rather evenly distributed above and below it – which makes it plausible to assume that the distribution resulting from larger samples would tend to be normal.

What is perhaps interesting to remark on is that variability in the frequency of initiating moves was higher than that for turn-exchanges, especially in the more interactive tasks. Although it is premature to draw any general conclusions, it would be worth investigating whether this trend holds for other contexts as well. If this proved to be the case, one might conclude that turn-taking dynamics is more inherent to the nature of each task - for which there is a 'physiological' rhythm of taking and yielding turns - while the proportion of initiatives may be more related to personal style and preferences, and thus more subject to inter-individual variation.

COMPUTING THE INTERACTIONAL DIFFICULTY INDEX

The two dimensions discussed above, number of turn exchanges and initiating moves per 100 words, together with visual access, have been argued to contribute to a task's interactional difficulty. The question is whether these dimensions can be combined in order to arrive at a global measure of interactional difficulty. The scales for the two dimensions had different ranges, so that their absolute values cannot be directly summed, but have to be standardized in order to become comparable. Given the small number of tasks and participants, it seems advisable to opt for a simple form of standardization making no assumptions on data distribution, which is ranking the observations into four quartiles. In order to do so, values by all learners in all tasks were cumulated into two scales of 60 observations (6 tasks * 10 participants), one for turn exchanges and another for initiating moves, and these were divided into quartiles, whose cut-off points are reported in Table 3.

[TABLE 3 NEAR HERE]

The difficulty index for each task was then calculated by assigning a score of 1 if that task's median value fell within the first, lower quartile, 2 if it fell in the second, and so forth.

As regards the eye-contact dimension, 2 points were given in its absence, and 0 points when participants could see each other. The choice of this particular score, which implies some degree of arbitrariness, can be argued on the ground that a 2-point difference represents a rise from one quartile to two above it, i.e. from a low to a medium-high difficulty or from a medium-low to a high difficulty, which seems to be a reasonable increment produced by this additional parameter.

These three criteria thus produce three scores that are added to arrive at a global interactional difficulty index (IntD) for each task (Table 4).

[TABLE 4 NEAR HERE]

Quite expectedly, the film retelling has the lowest IntD score, being essentially a monologue with very few chances for participants to exchange turns or take initiatives. It is followed by the

interview, in which the focal participant produces shorter turns (though not very short, compared with other tasks), but never or almost never takes an initiating role. The decision-making negotiation has a brisk turn-taking pace (fourth quartile), but does not require to take the initiative very often, as this role can also be played by the other participant or by the adult. Telephone calls offer a complementary picture, with fewer turn exchanges due to the rather long answers provided by shops, offices or experts, but with the focal participant playing a more active role in asking for information and advice. In this task there is also no eye-contact among participants, which introduces a further aspect of interactional difficulty. There is no eye-contact in map tasks either, but turn-taking here is more lively (third and fourth quartile for map tasks with adult and peer, respectively) and the focal participant must take the initiative more frequently (fourth quartile for both tasks).

The six tasks may thus be ordered on a scale of increasing interactional difficulty, ranging from the film retelling's minimum to a maximum in the map task with peer. The scale is empirically grounded, as its values derive from the observation of native speakers' task performance, and cut-off points between different scores on each dimension are identified in an objective way, by dividing the whole distribution of observed values into quartiles.

CONCLUSIONS

This article made some methodological proposals contributing to the current debate on how task difficulty should be conceptualized and measured, which is a key issue for SLA research, Task-Based Language Teaching and Task-Based Language Assessment.

Firstly, a particular dimension of task difficulty, interactional difficulty, has been identified and set apart from other factors, such as the lexical and grammatical structures required (linguistic difficulty) or the range of speech acts and politeness constraints (pragmatic difficulty). Interaction as an independent variable has received little attention thus far in SLA research on tasks, and the few existing studies have operationalized it in dichotomous terms (± interaction). We have instead shown that different types of dialogic exchange – such as an interview, a telephone call, a discussion or a map task – may be quite different in terms of their interactional dynamics.

Secondly, a new procedure has been proposed, which operationalizes difficulty essentially in terms of the communicative skills required by different tasks. The procedure consists in the following steps: 1) identifying difficult communicative behaviors based on how long it takes learners to master them; 2) observing native speakers' performance to see to what extent they tend to produce these behaviors in different tasks, which can thus be said to require, by their very nature, the performance of difficult communicative actions. These observations allow one to construct a difficulty scale that is more fine-grained than a '± complex' dichotomy and whose cut-off points are empirically grounded, rather than involving subjective judgments about vague notions like *few*, *many*, *more*, *less*.

It is possible to follow the same procedure to gauge other aspects of task difficulty, for example lexical difficulty. In this case, one would observe native speakers' productions in order to assess the presence of features that have been shown to require a longer learning time in L2 acquisition, such as lexical diversity or a high proportion of rare words (Milton 2009). Tasks may then be ordered according to whether they fall in a high or low quartile with respect to lexical diversity and/or sophistication, and these two dimensions may be combined to provide a global lexical difficulty index. A similar approach may be followed to evaluate further aspects such as grammatical difficulty (as defined e.g. by Housen and Simoens 2016) or pragmatic difficulty (as operationalized e.g. by Brown, Hudson, Norris, Bonk 2002; Taguchi 2007).

According to the two basic dictionary meanings of difficulty, this procedure can be said to essentially concern the dimension of *skill* - tasks are classified based on the type of communicative skills they involve or require. This approach thus continues a tradition of assessing task difficulty by

looking at task-takers' performance, with a significant methodological difference – rather than considering learners' performance, which is conditioned by their proficiency level, we have proposed to examine native speakers' performance. This of course is variable, too, as all things human, but the proficiency component can be considered to be (almost) constant, and performance variance can thus be attributed mainly to task characteristics, plus an inevitable quota of inter-individual variability. This way of viewing difficulty in terms of the skills involved by tasks is of particular relevance to language teachers and testers, who need principled ways for establishing and objectively measuring the linguistic-communicative demands of different activities and the skills they require.

The present approach to gauging task difficulty is to be seen as complementary to other approaches, such as those analyzing task-takers' difficulty perceptions, their subjective estimations of time on task, performance on a secondary task or physiological measures. All these operationalizations involve the other main sense of difficulty, namely *effort*. There is clearly a relationship between skill and effort – it is for instance well-known that automated behaviors, which have taken a long time to be acquired, require fewer cognitive resources (De Keyser 2014). Future research will have to investigate how these relationships can be theorized and assessed in the specific domain of task difficulty in SLA.

Collecting all this information for each task used in research, teaching and assessmentis clearly a tall order, so any single study will be limited to some extent. As regards the present study, its main limitation is that the sample of ten participants does not allow one to make reliable inferences about median and quartiles cut-off values for a larger population of *Italian female teenagers performing these six tasks*, let alone for a population of persons of various ages and from different backgrounds. Its main contribution is thus conceptual and methodological, as an attempt to propose a new approach to assessing task difficulty, exemplified on one specific dimension, interactional difficulty. Still, even with these limitations, the study involved collecting some 20 hours of video-recorded interactions and transcribing and analyzing over 70,000 words. Given the high costs associated with task-based research, Long (2015: 239) concludes that, 'rather than more and more one-off studies using a miscellany of variables, measures, and analyses, what is needed is a unified *research program*, albeit conducted by individual researchers and research groups'.

Such a research program might be implemented through a Collaborative Research Network, like the one sponsored by AILA (www.aila.info/en/research/list-of-rens/tasks-and-second-language-learning.html), whose goal is 'to promote consistency in the way in which task-based research operationalizes independent and dependent variables'. In the current state of affairs, each study, including ours, can only say whether a task is more difficult than another among the few taken into consideration in that particular research project. The scientific community should now strive to address 'big problems' by cooperatively working on a large scale (Norris 2016), which would mean gathering multiple sources of information about a few well-studied tasks, that may then become standards for conducting systematic investigations on the relationships between task characteristics and L2 acquisition and use (see e.g. Révész, Michel and Gilabert 2016). Following the present operationalization, for instance, if a substantial number of data points were available for several tasks, each performed by many participants, the relative difficulty of any new task could be established by reference to the quartiles of this large distribution, with a much higher degree of reliability and generalizability than can be afforded by any individual study, whose conclusions will always be limited to the small set of tasks taken into consideration.

ACKNOWLEDGMENTS

I wish to thank Ulrike Arras, Paolo Della Putta, Alex Housen, Paola Leone and the anonymous reviewers for their comments on previous versions of this article. All remaining errors are my own.

REFERENCES

Baralt, M., Gilabert, R. and **Robinson, P.** 2014. 'An introduction to theory and research in task sequencing and instructed second language learning,' in M. Baralt, R. Gilabert, and P. Robinson (eds.): *Task Sequencing and Instructed Second Language Learning*. Bloomsbury.

Brooks, L. 2009. 'Interacting in pairs in a test of oral proficiency: Co-constructing a better performance,' *Language Testing*, *26*(3): 341-366.

Brown, J. D., Hudson, T., Norris, J., Bonk, W. 2002. *An investigation of second language task-based performance assessments*. University of Hawaii Press.

Campbell, D. J. 1988. 'Task complexity: A review and analysis,' *Academy of Management Review 13(1):* 40-52.

Cekaite, A. 2007. 'A Child's Development of Interactional Competence in a Swedish L2 Classroom,' *The Modern Language Journal 91(1)*: 45-62.

Dahl, T. I. and **Ludvigsen, S.** 2014. 'How I See what you're saying: The role of gestures in native and foreign language listening comprehension,' *The Modern Language Journal*, *98*(3): 813-833.

DeKeyser, R. 2014. 'Skill acquisition theory,' in B. VanPatten and J. Williams (eds.): *Theories in Second Language Acquisition, 2nd ed.* Routledge.

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. 2003. *Task-based language learning and teaching*. Oxford University Press.

Foster, P. and **Tavakoli, P.** 2009. 'Native speakers and task performance: comparing effects on complexity, fluency, and lexical diversity,' *Language Learning, 59(4)*: 866-896.

Fulcher, G and **Márquez Reiter, R.** 2003. ' Task Difficulty in Speaking Test,' *Language Testing*, *20(3)*: 321-344.

Gilabert, R., Barón, J. and **Levkina, M.** 2011. 'Manipulating task complexity across task types and modes,' in P. Robinson (ed.): *Second Language Task Complexity. Researching the Cognition Hypothesis of language learning and performance.* Benjamins.

Gilabert, R., Barón, J. and **Llanes, A.** 2009. 'Manipulating cognitive complexity across task types and its impact on learners' interaction during oral performance,' *International Review of Applied Linguistics in Language Teaching, 47(3-4)*: 365-395.

Gill, T. G. and **Hicks, R. C.** 2006. 'Task complexity and informing science: A synthesis,' *Informing Science, 9:* 1-30.

Granena, G. and **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.

Goffman, E. 1981. *Forms of talk*. University of Pennsylvania Press.

Gullberg, M. 1998. *Gesture as a communication strategy in second language discourse: A study of learners of French and Swedish*. Lund University Press.

Gullberg, M. 2006. 'Handling Discourse: Gestures, Reference Tracking, and Communication Strategies in Early L2.' *Language Learning 56(1):* 155–196.

Hall, J. K., and **Pekarek Doehler, S**. 2011. 'L2 interactional competence and development' in J. K. Hall, J. Hellermann and S. Pekarek Doehler (eds): *L2 interactional competence and development*. Multilingual Matters.**Housen, A.** and **Simoens, A.** 2016. 'Introduction: cognitive perspectives on difficulty and complexity in L2 acquisition,' *Studies in Second Language Acquisition*, 38: 163 – 175.

Hulstijn, J. 2015. *Language proficiency in native and non-native speakers: Theory and research*. Benjamins.

Kane, M. T. 2013. 'Validating the interpretations and uses of test scores.' *Journal of Educational*

Measurement, *50*(1): 1-73.
Kim, Y. 2009. 'The effects of task complexity on learner-learner interaction.' *System, 37(2)*: 254-268.
Kuiken, F. and **Vedder, I.** 2007. 'Task complexity and measures of linguistic performance in L2 writing,'I*nternational Review of Applied Linguistics in Language Teaching,* 45: 261-284.
Long, M. 2015. *Second language acquisition and task-based language teaching*. Wiley-Blackwell.
Mackey, A. 2012. *Input,interaction and corrective feedback in L2 classrooms*. Oxford University Press.
Malicka, A. and **Levkina, M.** 2012. 'Measuring task complexity: does L2 proficiency matter?' in A. Shehadeh and C. Coombe (eds.): *Task-based language teaching in foreign language contexts: Research and implementation*. Benjamins.
Michel, M. C. 2011. 'Effects of task complexity and interaction on L2 performance.' in P. Robinson (ed.): *Second Language Task Complexity. Researching the Cognition Hypothesis of language learning and performance*. Benjamins.
Michel, M. C., Kuiken F. and **Vedder I.** 2007. 'The influence of complexity in monologic versus dialogic tasks in Dutch L2,' *International Review of Applied Linguistics 45(3)*: 241–259.
Michel, M. C., Révész, A. and **Gilabert, R.** (2014, Aug). Eye movement prompts in stimulated recall: tapping cognitive processes based on audio vs. visual stimuli. Paper presented at AILA, Brisbane (Australia).
Milton, J. 2009. *Measuring second language vocabulary acquisition*. Bristol: Multilingual Matters.
Nguyen, H.T. 2011. 'A longitudinal microanalysis of a second language learner's participation' in G. Pallotti, and J. Wagner (eds.), *L2 learning as social practice: Conversation-analytic perspectives (pp. 17-44.* National Foreign Language Resource Center.
Norris, J. M. (2016, June). Reframing the SLA-Assessment Interface: 'Constructive' Deliberations at the Nexus of Interpretations, Contexts, and Consequences. Invited Alan Davies Lecture at the Language Testing Research Colloquium, Palermo, Italy.
Nuzzo, E. and **Gauci P.** 2012. *Insegnare la pragmatica in italiano L2* [Teaching pragmatics in L2 Italian]. Carocci.
Pallotti, G. 2001. 'External appropriations as a participation strategy in intercultural multi-party interactions' in A. Di Luzio, S. Guenthner, and F. Orletti (eds.): *Culture in Interaction*, Benjamins.
Pallotti, G. 2009. 'CAF: Defining, refining and differentiating constructs,' *Applied Linguistics*, *30*(4): 590-601.
Pallotti, G. and **Ferrari S.** 2008. 'La variabilità situazionale dell'interlingua: implicazioni per la ricerca acquisizionale e il testing linguistico [Interlanguage situational variability: implications for SLA research and language testing]' in G. Bernini, L. Spreafico, and A. Valentini (eds.), *Competenze lessicali e discorsive nell'acquisizione di lingue seconde*. Guerra.
Pallotti, G., **Ferrari, S**. and **Nuzzo, E**. 2011. 'A systematic procedure for assessing communicative competence' in W. Wiater and G. Videsott (eds), *New theoretical perspectives in multilingualism research*. Peter Lang.
Pekarek Doehler, S. and **E. Pochon-Berger**. 2015. The development of L2 interactional competence: evidence from turn-taking orgnization, sequence organization, repair organization and preference organization' in T. Cadierno, and S. Eskildsen (eds): *Usage-based Perspectives on Second Language Learning*. Mouton De Gruyter.
Révész, A . 2011. 'Task complexity, focus on L2 constructions, and individual differences: A classroom-based study.' *The Modern Language Journal*, 95 (Suppl. 1): 162 – 181 .
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.
Révész, A., Michel, M. and **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.

Révész, A., Sachs, R. and **Hama, M.** 2014. 'The effects of task complexity and input frequency on the acquisition of the past counterfactual construction through recasts,' *Language Learning*, 64(3): 615-650.

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*. Benjamins.

Rossano, F. 2013. 'Gaze in conversation' in J. Sidnell and T. Stivers (eds.): *The handbook of conversation analysis*. Wiley-Blackwell.

Sacks, H., Schegloff, E. A. and **Jefferson, G.** 1974. 'A simplest systematics for the organization of turn-taking for conversation,' *Language, 50(4)*: 696-735.

Sasayama, S. 2016. 'Is a "Complex" Task Really Complex? Validating the Assumption of Cognitive Task Complexity,' *The Modern Language Journal*, 100: 231–254

Sasayama, S., Malicka, A. and **Norris, J.** 2015. 'Primary challenges in cognitive task complexity research: Results of a comprehensive research synthesis,' Paper presented at the 6th Biennial international Conference on Task-Based Language Teaching (TBLT), Leuven, Belgium.

Schegloff, E. A. 2007. *Sequence organization in interaction: A primer in conversation analyis, Volume 1*. Cambridge University Press.

Schegloff, E. A. and **Sacks, H.** 1973. 'Opening up closings,' *Semiotica,7:* 289-327.

Shehadeh, A. 2012. 'Task-based language assessment: Components, development, and implementation,' in C. Coombe, B. O'Sullivan, S. Stoynoff (eds.): *The Cambridge guide to second language assessment*. Cambridge University Press.

Skehan, P. 1998. *A cognitive approach to language learning*. Oxford: Oxford University Press.

Skehan, P. and **Foster, P.** 2001. 'Cognition and tasks,' in P. Robinson (ed.), *Cognition and Second Language Learning*. Cambridge University Press.

Sueyoshi, A. and **Hardison, D. M.** 2005. 'The role of gestures and facial cues in second language listening comprehension,' *Language Learning, 55(4)*: 661–699.

Taguchi, N. 2007. 'Task difficulty in oral speech act production,' *Applied Linguistics*, *28(1)*: 113-135.

Tavakoli, P. 2009. 'Investigating task difficulty: learners' and teachers' perceptions.' *International Journal of Applied Linguistics, 19*(1): 1-25.

Wagner, E. 2010. 'The effect of the use of video texts on ESL listening test-taker performance,' *Language testing*, *27(4)*: 493–513.

Young, R. 2011. 'Interactional competence in language learning, teaching, and testing,' in E. Hinkel (ed.), *Handbook of research in second language teaching and learning* (Vol. 2). Routledge.

Young, R. and **Miller, E.** 2004. 'Learning as changing participation: Discourse roles in ESL writing conferences,' *The Modern Language Journal, 88(4)*: 519–535.

Zuengler, J. and **Bent, B.** 1991. 'Relative knowledge of content domain: An influence on native-non-native conversations', *Applied Linguistics*, 12(4): 397–415.

Tables

	Median	*Min*	*Max*	*IQR*	*Mean*	*SD*
Film	1.96	0.28	3.98	1.43	1.87	1.19
Interview	9.87	4.38	12.86	3.73	9.37	2.92
Calls	12.21	10.05	17.93	2.1	12.72	2.25
Map-adult	18.69	15.12	20.34	2.72	18.18	1.88
Map-peer	19.5	13.88	23.29	2.39	19.48	2.60
Negotiation	19.69	13.86	21.48	6.52	18.01	3.46

Table 1. Turn exchanges per 100 words

	Median	*Min*	*Max*	*IQR*	*Mean*	*SD*
Interview	0.05	0	0.59	0.34	0.18	0.23
Film	0.18	0	0.54	0.18	0.22	0.19
Negotiation	2.05	1.14	3.78	0.88	2.24	0.84
Calls	3.34	0.92	5.42	1.18	3.22	1.22
Map-adult	5.04	3.49	7.98	2.25	5.22	1.47
Map-peer	6.36	3.74	8.07	1.89	6.23	1.46

Table 2. Initiating moves per 100 words

	Turns/100w	Initiatives/100w
0%	0.28	0.0
25%	9.79	0.38
50%	13.87	2.66
75%	19.41	4.81
100%	23.29	8.07

Table 3. Quartile cut-off points

	turns/100w	init/100w	visual access	**IntD score**
Film	1	1	0	**2**
Interview	2	1	0	**3**
Negotiation	4	2	0	**6**
Calls	2	3	2	**7**
MapAd	3	4	2	**9**
MapPeer	4	4	2	**10**

Table 4. Interactional Difficulty scores for different tasks

Figures

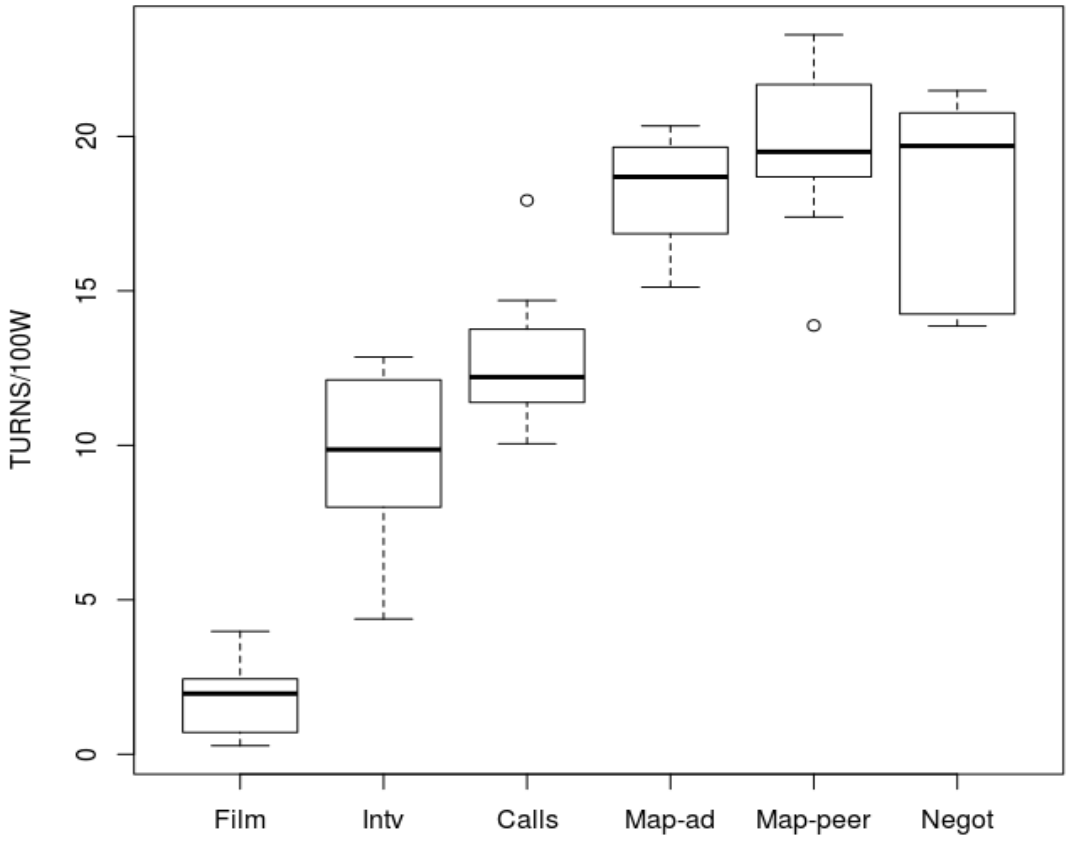

Figure 1. Turn exchanges per 100 words

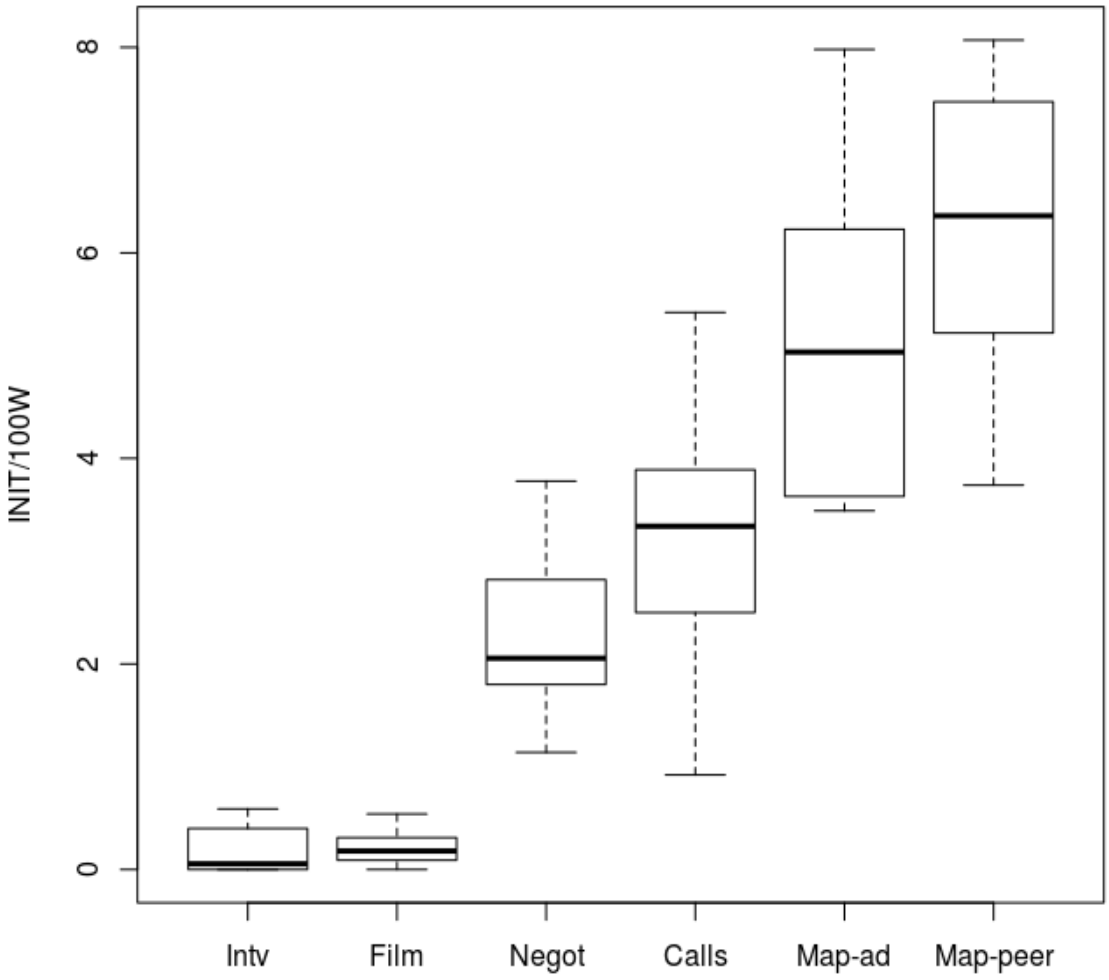

Figure 2. Initiating moves per 100 words