An exploratory study of second language oral performance: Fluency and lexical complexity in L1 Turkish and L2 English

Zeynep Duran-Karaoz

Thesis submitted for the degree of Doctor of Philosophy in Applied Linguistics

Department of English Language and Applied Linguistics

School of Literature and Languages

University of Reading

November 2020

Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Zeynep Duran-Karaoz

Acknowledgement

I would have not been able to complete this research without the support of many people around me. First of all, I would like to express my sincere gratitude to my supervisor Professor Parvaneh Tavakoli for her dedicated support and guidance, and ultimately her belief in me. Parvaneh continuously provided encouragement and was always willing and enthusiastic to assist me in any way she could throughout this research project. We have had many interesting, enjoyable and fruitful discussions together; I am eternally grateful to her. I would also like to thank to my second supervisor Dr. Michael Daller for being really supportive and providing helpful advice as this project developed. I have benefitted from his guidance and insight on so many levels. Thank you also to all the staff in Department of English Language and Applied Linguistics (DELAL), especially Dr. Jacqueline Laws, who had been the Director of Postgraduate Research Studies until recently, and Professor Rodney Jones, the Head of DELAL, for their encouragement and endless support. I have really enjoyed my time in the department as a PhD student, and the feedback and friendly support I have received in DELAL has made my PhD journey memorable.

I am indebted to the support of my friends and colleagues in Turkey, especially those in the International Relations Office at Dumlupinar University. I would like to mention particularly my best friends, Fatma Gelip Demirezen and Semra Onay Tas, who not only helped me in the school where the data collection of my first study took place but offered moral support in any other way they could during all these years. I have never doubted that they were perfectly sincere. Also, Ozge Cizmeci, who assisted me in reaching out to my participants and who bore with my endless requests in running this research, and all the students who voluntarily participated in my research both in Turkey and in the UK and enabled this project to be possible- thank you all so much!

Having had to be away from my country and my beloved ones for more than four years now has been very difficult and even unbearable at times; so I have particularly appreciated the support of my friends in the UK throughout these difficult times. Very special thanks to my dear friends Nihal and Ilkay for being there for me and my husband whenever we needed them, and to Anne, John, Sam and Ben for being like a family to us for years. You will always be remembered with gratitude.

I would like to say how much I have been supported by my amazing family and my husband's family. It is not possible to mention everyone individually here, but I would like to thank in particular to my parents (Sündüs and Mustafa), my sisters (Dilek, Melek and Tugba), and to my nieces and nephews (Zehra Naz, Elif Beyza, Ahmet Salih, Elif and Mustafa) whom I apologise for not being there with them as I should have been over the last five years. And finally, very special thanks to my favourite person in my life, my husband Ismail who has recently finished his PhD education. Ismail, I owe so much to you- thank you for putting up with me, sacrificing so many things for me and believing in me. Without you, I would have not been able to make it. We both have finished now. Are you ready for a new adventure with me?

Table of Contents

List of Tables

List of Figures

List of Abbreviations

BWDS	Backward Digit Span Test
CAF	Complexity, Accuracy, and Fluency
CALF	Complexity, Accuracy, Lexis and Fluency
CEFR	Common European Framework of Reference for Languages
ECFP	End-clause Filled Pauses
ECSP	End-clause Silent Pauses
EFL	English as a Foreign Language
EIT	Elicited Imitation Task
H	Hypothesis
L2	A Foreign or Second Language
LD	Lexical Diversity
LoR	Length of Residence
LP	Language Proficiency
LS	Lexical Sophistication
LTM	Long-term Memory
MCFP	Mid-Clause Filled Pauses

MCSP	Mid-Clause Silent Pauses
OOPT	Oxford Online Placement Test
OPT	Oxford Placement Test
PWM	Phonological Working Memory
RQ	Research Question
SNRT	Serial Non-Word Recognition Task
TL	Target Language
TTR	Type/Token Ratio
WM	Working Memory
WMC	Working Memory Capacity

Abstract

Understanding second language (L2) fluency has been an area of central importance for L2 research. While it is generally assumed that fluency improves with increased L2 proficiency, little attention has been paid to the effects of non-L2 specific features of oral performance, such as personal speaking styles in first language (L1). This thesis draws on the findings of two separate but interrelated studies. Study 1 aims to explore the relationship between L1 and L2 fluency behaviours amongst lower proficiency learners (A2, B1 and B2 at CEFR) in an English as a Foreign Language (EFL) context. It also addresses the question of whether the relationships, if any, are mediated by variations in L2 proficiency and task structure. The data were collected from 42 Turkish learners of English who were undergraduate students at a university in Turkey. Language proficiency was measured through two standardized tests of Oxford Placement Test (OPT) and Elicited Imitation Task (EIT), and speech samples were elicited through two structurally different narratives, i.e. tight and loose structure. The data were coded for 17 fluency measures; i.e. eight breakdown, five repair, one speed and three composite measures. The results from the correlational analysis showed that some breakdown and repair measures were positively correlated in L1 and L2, but no correlations were found for speed and composite measures. The relationships were not mediated by variations in L2 proficiency and the task structure. The regression analysis demonstrated a number of models predicting L2 fluency. While L1 fluency contributed significantly to models predicting breakdown and repair fluency, the OPT scores predicted mid-clause pausing and reformulation and the EIT scores predicted speed-related composite measures of speech rate, phonation-time ratio and mean length of run.

Study 2 aims to further examine L1-L2 fluency relationships in a study-abroad context amongst learners of higher proficiency levels (B1, B2 and C1 according to CEFR). It also explores the mediating roles of three individual learner variables on these links; i.e. L2 proficiency (measured through the OPT and the EIT), working memory capacity (WMC) (measured through Backward digit span test and serial recognition task) and length of residence (LoR) (represented through the amount of time spent studying abroad). 60 Turkish learners of English who were post-graduate students in the UK participated in

Study 2. They performed two oral narrative tasks and their speech were coded and analysed for five fluency measures (two breakdown, one repair, one speed and one composite). The correlational analysis revealed positive correlations between L1 and L2 fluency measures. They also showed that the correlations were overall maintained when the effects of L2 proficiency, WMC and LoR were controlled for. Study 2 was also interested in examining the link between L1 and L2 lexical complexity behaviours and between fluency and lexical complexity aspects in each language separately. Lexical complexity was analysed in terms of lexical diversity (LD) using two measures of type/token ratio (TTR) values and D scores. The findings revealed that LD in L1 and L2 were related to a small extent when measured through TTR values. Weak correlations were observed between fluency and LD in each language, and only between speed fluency and LD measures. When taken together, the results imply that L1 personal speaking styles are carried over to L2 fluency behaviour at least to some extent, and to L2 lexical complexity to a limited extent. The findings have a number of significant implications for L2 research, testing and teaching practices.

CHAPTER 1. INTRODUCTION

1.1. Introduction

Speaking a foreign or second language (L2) effortlessly has been a primary goal for many language learners. While most learners can build a substantial knowledge about the language (e.g. grammar, rules vocabulary or syntax), it is their oral communication ability that they find hard to improve, even after studying the language for years. Clearly, successful communication is more than knowing the grammar or vocabulary but involves being able to implement the linguistics knowledge to real time performances effectively and effortlessly. In other words, learners need to translate what they intend to say into comprehensible speech rapidly and without disrupting the flow of speech (e.g. with fewer pauses or hesitations) under the constraints of time. This ability is often referred to as 'L2 fluency' by language researchers.

Developing a good understanding of L2 fluency, including how it develops and what factors affect its development is important for several reasons. First, fluency is featured as one of the assessment criteria for oral proficiency assessment in several high-stake international language tests (e.g. APTIS, IELTS) and language benchmarks (e.g. CEFR, 2001). Often, test results would significantly inform many important decisions in peoples' lives such as those about their education, employment or immigration. This means that learners become even more motivated towards improving their fluency to obtain better results from their exams/tests. Second, fluency has a pivotal place in teaching practices as well whether in different instructional contexts such as schools, private language courses or universities or different learning contexts such as English as a Foreign Language (EFL) or study-abroad programmes. In language classrooms, it is common to observe that learners' overall progress in improving their interlanguage is tested against their fluency performance. In fact, some L2 teachers may even equate fluency performance with overall language proficiency (LP) (Tavakoli & Hunter, 2018).

Understanding fluency has also been a key concern for L2 researchers. In SLA field, fluency is seen as a phenomenon that is revealing different processes underlying speech production (Bosker, Pinget, Quené, Sanders, & De Jong, 2013; De Jong, 2016; De Jong, Steinel, Florijn, Rob Schoonen, & Hulstijn, 2012b; De Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2013; Segalowitz, 2010, 2016). A number of studies can be found in the literature in this regard; i.e. studies examining temporal features of speech (e.g. frequency of pauses or corrections) aiming to help develop a more in-depth understanding of underlying cognitive processes in speech (Baddeley, 1974; Baddeley, Hitch, & Allen, 2009; De Jong, 2016; Tavakoli, 2011). The construct of fluency is also suggested to be one of the key factors in predicting overall L2 proficiency (Baker-Smemoe, Dewey, Bown, & Martinsen, 2014; Iwashita, Brown, Mcnamara, & O'Hagan, 2008; Révész, Ekiert, & Torgersen, 2016). In these studies, fluency has emerged as a critical component of speaking proficiency (Ginther, Dimova, & Yang, 2010) or communicative adequacy (Révész et al., 2016). When taken together, these points indicate that fluency has been an important concept in SLA practices whether for research, teaching or testing purposes.

Yet, fluency has been difficult to pin down in L2 research. Most researchers examining oral fluency are likely to agree that 1) fluency is a construct which is difficult to conceptualize and measure, and 2) a number of social and psychological factors can impact its development and performance (Segalowitz, 2016). While fluency can be a subjective phenomenon on the part of the listener residing in the ear and mind (Freed, 2000), it can also be an objective phenomenon related to underlying cognitive mechanisms which could be observed and measured (Segalowitz, 2010). When it comes to its measurement, there does not appear to be a single approach across studies on how to measure fluency; it is still not entirely clear which features of speech could be better or best fluency indicators. Another important question in this area is to what extent fluency analysis could be carried out automatically through the use of computer technologies. Although fluency analysis has historically relied on manual identification and counts of dysfluencies in speech (e.g. pauses, speech rate), the development of specialist software such as PRAAT has made it possible to detect some of such speech features (e.g. silent pauses) automatically, making fluency analysis more feasible, objective and precise (De Jong & Wempe, 2009). Yet, PRAAT is still far from a completely reliable tool which can detect all dysfluencies in speech automatically or can cope with large amounts of data.

A further crucial issue relates to understanding sources of influence on fluency development. L2 research has shown that a number of internal and external factors could potentially impact L2 fluency development/performance including, but not limited to, speakers' linguistic development (e.g. Cucchiarini, Strik & Boyes, 2000, 2002, Ginther et al., 2010; Iwashita et al., 2008), L1 fluency (e.g. De Jong, Steinel, Florijn, Schoonen & Hulstjin, 2012a; 2015; Derwing, Munro, Thomson, & Rossiter, 2009; Huensch & Tracy-Ventura, 2016), cognitive skills such working memory capacity (e.g. Kormos & Safar, 2008), task design features such as task structure (e.g. Skehan & Foster, 1996; Tavakoli & Foster, 2011) or language practice (e.g. Mora & Valls-Ferrer, 2012; Segalowitz & Freed, 2004). What is more, there seems to be a complex and interactive relationship among the factors that influence L2 fluency. While great progress has been made in L2 research in terms of conceptualizing fluency as a distinction is now made between different types of fluency (i.e. cognitive, perceived and utterance) (Segalowitz, 2010, 2016) and its measurement in line with the development of specialist software (De Jong & Wempe, 2009), our understanding of fluency behaviour and knowledge about different factors that have a bearing on it remains rather limited. Clearly, we need to extend our understanding in this regard by exploring sources of influence on L2 fluency performance or development.

1.2. Background to this research

Becoming a fluent speaker may be the most challenging skill for most L2 learners, especially in EFL contexts such as in my country, Turkey. This is largely because in EFL contexts, for example, English is mainly taught through formal interaction in classrooms and learners have fewer opportunities to engage in real-life L2 communication outside class. However, throughout the years that I taught English as a foreign language, I began to develop a different perspective on learners' fluency performance and the possible reasons behind seemingly disfluent speech. Observing my students' speech in their L1 and L2, I noticed certain characteristics of their speech (e.g. frequent uses of pauses, reformulating what was being said or even the lexical features such as the use of high or low-frequency words) had an impact on how fluent they came across. I was then intrigued by this interesting question: ‘Could a person who exhibits frequent pauses in their L1 speech be expected to speak fast in their L2 speech?’. Fluent speech is usually characterised by features such as faster speech rate or fewer hesitations, corrections or pauses and judgements can be made on L2 speakers’ fluency, based largely on such indices in their speech, whether in classrooms contexts, job interviews or language tests. However, it is known that native speakers similarly exhibit such features in their speech; indeed, L1 speakers do differ widely in how much they pause, how fast they speak or how linguistically complex (syntactic or lexical) speech they produce. If these features are part of peoples’ speaking styles in their L1, it seems plausible to expect the same features to surface in L2 speech as well.

Although the interest for this research was initially inspired by my personal observation, the motivation to look into the relationship between L1 and L2 oral performance with respect to fluency (mainly) and lexical complexity (partly) came from my reading of the literature; the existing research evidence suggests links between the two languages in these regards. Regarding speech fluency, research has raised an interesting question of whether fluency is a characteristic specific to an individual (i.e. trait-like) or a language (i.e. state-like) (De Jong, Groenhout, Schoonen, & Hulstijn, 2015; Derwing et al., 2009; Riazantseva, 2001; Segalowitz, 2010). Derwing et al. (2009) notes that if fluency is related to a language, one can expect a speaker to be highly fluent in one language while not so in another. Yet, fluency is also ‘a defining feature of an individual’s first language (L1) output and (therefore) one might expect it is a trait; i.e. a relatively permanent characteristic specific to an individual’ (p. 534). In the latter case, it would be highly likely that a fluent speaker in L1 would be fluent in L2 as well. Subsequently, our attention has been brought to the importance of understanding how much of L2 fluency behaviour could be attributed to L1 fluency behaviour. Although there have been few studies looking into this, the findings from recent studies (e.g. De Jong et al., 2015; Derwing et al., 2009; Huensch & Tracy-Ventura, 2016; Peltonen, 2018) seem to imply that L1 personal speaking styles could at least to some extent be carried over to L2 speech fluency. This means that L2 fluency as seen by most researchers could not be a purely

L2-specific phenomenon, but it is related to L1 fluency behaviour, i.e. speech features typically perceived as signs of disfluency (e.g. frequent pauses, corrections or slower speech rate) may reflect speakers' L1 styles. Regarding lexical complexity aspect, a review of the literature in this area also suggests that L2 lexical complexity could be sensitive to cross-linguistic effects (Daller, Van Hout, & Treffers-Daller, 2003; De Clercq, 2015; Dewaele & Pavlenko, 2003; Treffers-Daller, 2013) and stylistic variations (Laufer, 2003; Pallotti, 2015). Taken together, research in this area directed my attention to the need for a better conceptualization of both aspects of oral performance through a careful in-depth investigation of these in L1 and L2 speech.

1.3. Identifying gaps in the literature

Although previous research highlights the importance of exploring L1 effects on L2 fluency behaviour, cross-linguistic studies investigating this area/topic are scarce with only a few exceptions (De Jong et al., 2015; Derwing et al., 2009; Huensch & Tracy-Ventura, 2016; Peltonen, 2018). In this limited body of research, few languages have been examined so far with most studies focusing on typologically similar languages, e.g. English-French in Huensch and Tracy-Ventura (2016) or English- Spanish in De Jong and Mora (2019). The existing research suggests that typological differences and similarities between L1 and L2 could play a considerable role in the development of L2 fluency (Derwing et al., 2009; Huensch & Tracy-Ventura, 2016) in the sense that typologically closer language pairs are likely to help improve L2 fluency. This means that L1 effects could be reflected in different ways in L2 speech; i.e. through personal speaking styles or cross-linguistic effects. While an investigation of typologically similar languages could contribute to a deeper understanding of L1 effects in this regard, we need to extend our knowledge with further research by examining structurally distant language pairs. The current research aims to explore L1-L2 relationships between two typologically distant languages; i.e. L1 Turkish and L2 English.

LP involves the knowledge about a language and the ability to access and use that knowledge across different domains such as listening, speaking, reading or writing (Hulstijn, 2015). LP is known to have a crucial role in fluency development in the sense that an increase in LP is normally associated with more fluency. In other words, as learners expand their linguistic knowledge, their speech becomes more fluent. The findings from recent studies highlight that learners' fluency patterns do not change in a linear fashion across different LP levels (Tavakoli, Nakatsuhara, & Hunter, 2019) yet, no previous study, to the best of my knowledge, has explored whether L1- L2 fluency relationships persist across all levels. Equally important, in earlier studies LP has been investigated in a rather limited way. Most of these studies have focused on examining fluency behaviours of one or two LP groups of learners; i.e. intermediate to advanced groups in De Jong et al. (2015), De Jong and Mora (2019), Riazantseva (2001) or Towell, Hawkins and Bazergui (1996). In addition, LP has not been assessed systematically in these studies as most of them have used only one test, usually a grammar-based or a vocabulary test (De Jong et al., 2015). Although a grammar test might be a useful and practical way of testing learners' overall LP, relying on a grammar test to evaluate oral ability of the learners could be problematic mainly because a full picture of learners' LP may not be gained. A pen-and-paper test such as a vocabulary test might fall short of testing speaking skills reliably, or similarly, a spoken test such as Elicited Imitation Task (EIT) would be limited in testing knowledge about grammar. However, in order to be able to draw valid and reliable conclusions about LP's role in any fluency study, it is necessary to measure proficiency from a broader perspective by working with learners from more than two LP groups, including lower-level learners, and by measuring LP more systematically through tests that tap into different aspects of proficiency. Only then can a better understanding of L2 proficiency as a mediating factor on L1-L2 fluency relationships be reached.

Another individual learner factor contributing to variations in fluency performance is working memory capacity (WMC). Working memory is defined as a limited capacity system which is responsible for maintaining the information temporarily and processing it when speakers perform complex tasks such as L2 speaking (Baddeley, 2003, 2015, 2017). In L2 speech, the production process could even be more complex and demanding for the speaker depending upon LP level and degree of automaticity (Kormos, 2006). WMC is believed to play a significant role in this process, i.e. a greater capacity of working memory in individuals helps ease processing demands in L2 speech (e.g. by allocating more sources to different stages of speech production), leading to increased fluency. Therefore, learners with higher WMC are assumed to be at an advantage in fluency performance. This suggests that WMC as an individual learner factor could exert a potential impact on the relationship between L1 and L2 fluency behaviours; yet, no previous study has investigated this.

The amount of L2 experience is suggested to be another factor in explaining variations in L2 oral performance (Saito, 2013; Saito & Brajot, 2013). Often referred to as LoR, L2 experience has mostly been examined in study-abroad contexts (Di Silvio, Diao, & Donovan, 2016; Huensch & Tracy-Ventura, 2017). In these contexts, learners are assumed to have more opportunities to be engaged in authentic communication situations and be exposed to different aspects of language on an everyday basis. There is abundant research evidence (Díaz-Campos, 2004; Munro & Derwing, 2008; Saito & Brajot, 2013) to suggest that LoR is most beneficial for the development of speaking abilities in general and for L2 fluency more specifically (Du, 2013; Huensch & Tracy-Ventura, 2017; Kim et al., 2015; Mora & Valls-Ferrer, 2012). Similar to the role of other individual learner variables, i.e. LP and WMC, a review of the literature reveals that little is known about whether LoR mediates L1-L2 fluency links.

In addition to the individual learner factors mentioned above, there are also external factors which could potentially influence L1-L2 fluency relationships. One such factor relates to task design in terms of its degree of structure. Task structure refers to the underlying macrostructure of a task. For example, in a picture-based narrative task, task structure pertains to the extent the picture prompts in a task are closely connected to each other; a tight structure means that the prompts make up a story in a sequenced order (with a clear beginning, middle and an end) while a loose structure means the order of the prompts could be rearranged or changed without compromising the story (De Jong & Vercellotti, 2016; Tavakoli & Foster, 2011). It is assumed that when learners perform a loosely structured tasks, they will be required to engage in more cognitive processing (e.g. by creating meaningful links between the prompts) whereas in a tightly structured task, the processing demands would be eased off because the sequenced order of the prompts can be easily seen and understood. In studies which tested these assumptions, it was reported that tight structure leads to greater fluency in L2 speech (Foster & Tavakoli, 2009; Skehan & Foster, 1999; Tavakoli, 2009; Tavakoli & Foster, 2011; Tavakoli & Skehan, 2005). It might follow that task structure could also have an impact on the strength of L1-L2 relationships; yet, the question of whether task structure has a mediating role in this regard remains unanswered.

It is also important to note that cross-linguistic research examining L1 and L2 speech fluency has to a large extent been performed with learners in a study-abroad context or participants immersed in the target language (TL) community (De Jong et al., 2015; Derwing et al., 2009; Huensch & Tracy-Ventura, 2016; Riazantseva, 2001; Towell et al., 1996) with few studies situated in an EFL context where L2 learners have limited contact with the TL community or use L2 for everyday communication. This is a shortcoming in fluency research because, as was mentioned earlier, fluency development is an aspect of speaking ability which is most likely to benefit from L2 experience in study-abroad contexts. Therefore, in order to have an in-depth understanding of the relationship between L1 and L2 fluency, we need to explore fluency in different learning contexts including EFL contexts.

Finally, although the present study is primarily interested in examining the effects of L1 styles on L2 fluency, it also aims to examine lexical complexity, which is another important aspect of L2 oral performance, between the two languages. Similar to the examination of fluency aspect, the motivation for examining lexical complexity in the present research came from both my personal observations and the suggestions of existing research in this area. Lexical complexity is suggested to be another key factor closely related with L2 proficiency (Iwashita et al., 2008; Révész et al., 2016; Zareva, Schwanenflugel, & Nikolova, 2005). The findings from several studies have implicated cross-linguistic effects on lexical complexity development (De Clercq, 2015; Marsden & David, 2008) and performance (Daller et al., 2003). L1 effects could be observed through cross-linguistic effects and stylistic variations. It could be different in two structurally different languages; i.e. Turkish and English in the current case. For example, the number of function words in Turkish would be lower than that in English because Turkish is a highly agglutinative language (i.e. one that primarily uses agglutination and uses fewer but longer words than English to convey the same content due to the agglutinating morphology) (Oflazer, 2014), and function words are mostly embedded in morphemes (e.g. ‘evinde’, meaning ‘in his/her house’ in English). Lexical complexity is also sensitive to stylistic variations in L1 speech (Pallotti, 2015). Following a similar line of reasoning for speech fluency, it could be argued that L1 speakers do vary in the lexically complexity of their speech (e.g. speech with a higher percentage of low-frequency words), and their lexical complexity styles (e.g. tendency to speak with high-frequency words) could be reflected in their L2 speech. However, there has been little research exploring the effects of L1 styles in L2 lexical complexity. More research is needed to investigate this relationship, especially in typologically distant languages such as L1 Turkish and L2 English.

1.4. The present study

The present study mainly seeks to answer the calls of previous research (De Jong et al., 2015; Derwing et al., 2009; Segalowitz, 2010) by addressing the question of whether L2 fluency is a characteristic specific to an individual and a language. As was discussed above, one way to do this is to carefully examine fluency behaviour in L1 and L2 speech and understand to what extent variability in L2 fluency behaviour could be explained by L1 fluency.

As such, the main purpose of this research is to explore whether there is a relationship between fluency behaviours in L1 Turkish and L2 English. The research is designed in a way that aims to help fill the gaps outlined above and extend our knowledge of fluency behaviour and speech processing in L1 and L2. As part of this research, two separate but interrelated studies are carried out; i.e. Study 1 and Study 2. While the overriding focus of both studies is on examining L1-L2 speech fluency links, several differences exist between the two which make original contributions to the research in the area of fluency.

Situated in an EFL context, Study 1 investigates fluency links between L1 and L2 including learners of low LP. In terms of the number and range of measures employed (17 in total), it is an exploratory study that is aimed to provide an in-depth insight into which fluency measures correlate in L1 and L2. Study 2, on the other hand, is carried out in a study-abroad context and primarily investigates fluency behaviour for higher-level learners. While Study 2 build up on Study 1's findings as it also seeks to explore L1-L2 fluency links, a vital part of Study 2 is also interested in examining L1 and L2 lexical complexity behaviours, and the links between fluency and lexical complexity in each language separately. An important methodological aspect of both studies is that LP is examined from a broader perspective than in most previous studies. This is achieved in two aspects; 1) both studies examine fluency behaviours of three groups of language proficiency as opposed to one or two groups in most other studies, namely, lower-levels of A2, B1 and B2 at Common European Framework of Reference for Languages (CEFR) in Study 1 and higher-levels of B1, B2 and C1 at CEFR in Study 2, and 2) L2 proficiency is assessed systematically in both through the use of two standardized proficiency tests, i.e. Oxford Placement Test (OPT) and Elicited Imitation Task (EIT). Additionally, in both Study 1 and Study 2 each of the proficiency tests is intended to measure a different aspect of L2 proficiency; i.e. OPT to test declarative knowledge (mostly) and EIT to test procedural knowledge (mostly).

The present study is also motivated to find out whether variations in three individual learner variables, namely, L2 proficiency, WMC and LoR, and one external variable of task structure mediate L1-L2 fluency relationships. The mediating role of L2 proficiency on L1-L2 fluency links is studied in both Study 1 and Study 2, while the role of task structure is explored in Study 1 only, and WMC and LoR in Study 2. As the above discussion also indicates, these variables have been chosen in the current research as each of them have been found to have potential impacts on fluency development and/or performance, and each of these is therefore likely to mediate the possible L1-L2 fluency links either by hindering or facilitating it.

1.5. The layout of the thesis

The overall structure of this thesis takes the form of nine chapters. In this chapter, I have already provided an account of why it is important to understand and investigate L2 fluency. I also have explained the motivations for this research, presented a discussion of the gaps in the literature, and provided an overview of the two studies reported in this thesis.

In Chapter 2, I present a review of the relevant literature. Drawing on most widely referenced psycholinguistic models of speech production, I explore the similarities and differences between L1 and L2 speech production models. Then, I provide a detailed discussion of two important aspects of second language oral performance under two main sections, i.e. fluency and lexical complexity. Allocating a larger space for the fluency aspect, I look at some of the existing definitions of fluency in the literature and a working definition in the present research. Under the fluency section, I move on to discuss different aspects of fluency, some of the approaches previous studies have taken to examine fluency, and the fluency measures that L2 researchers have identified and used as reliable indices of fluency. These then lead up to an overview of cross-linguistic studies examining L1 and L2 speech fluency; I discuss the shortcomings of these studies and point at implications of the findings from these for the current research. Under the lexical complexity section, I briefly explain what is meant by word knowledge and highlight the multi-dimensional nature of lexical knowledge. I provide a detailed account of how different aspects of lexical complexity have been defined and studied by L2 researchers. Last section of this chapter presents a discussion of the independent variables of the present research (i.e. three individual learner factors of L2 proficiency, WMC and LoR and one external variable of task structure) as well as their important roles in fluency development and performance. In the light of previous studies, I also explain how variations in each of these factors may impact the relationships between L1 and L2 fluency behaviours.

The following three chapters are devoted to Study 1. In Chapter 3, I present the research questions (RQ) of Study 1 and the corresponding hypothesis (H) that guide this study. Introducing research design, information about the participants and context of research, I provide an explanation of the methodological approach (i.e. quantitative approach) used in Study 1 together with the underlying rationales for this choice. I then move on to give a detailed account of the rationale for the choice of research materials used and the procedures followed in data collection. This chapter also provides an in-depth description of data analysis procedures including data transcription and coding, calculation of fluency measures and interrater reliability. In Chapter 4, I describe the statistical analysis, i.e. Pearson product-moment correlation analysis, partial correlations and multiple regressions which were carried out to answer the RQs and provide justifications for using them in the study. Then, I present the results obtained through these. Chapter 5 is the final chapter dedicated to Study 1. In this chapter, first the key findings of Study 1 are summarised, and then the results are compared with those reported in previous studies. This chapter concludes with implications for Study 2.

Chapter 6 is concerned with the methodology used in Study 2. I start with RQs and hypotheses guiding the study and move on to outline research design and instruments used, providing justifications for my choices. Since the procedures followed in data collection and analysis were similar to those in Study 1, my aim in this chapter is to focus on discussing the challenges faced in Study 2 regarding these stages and describing the steps that are specific to this study. This chapter also presents detailed information about how the data were analysed for lexical complexity, measures used, and the challenges faced in lexical analysis. While Chapter 7 explains statistical analysis carried out to address the RQs of Study 2 as well as the results obtained through these, Chapter 8 discusses these findings in the light of the existing literature and compares them with those reported in Study 1.

Finally, in Chapter 9, I first provide a summary of Study 1 and Study 2 and highlight the key findings. I then explain how this thesis contributed to our existing knowledge and the research in this area, and discuss its theoretical, methodological and pedagogical implications. I will then discuss the limitations of this research and provide directions for future research.

CHAPTER 2. LITERATURE REVIEW

2.1. Introduction

Fluency and lexical complexity have been two important aspects of L2 oral performance that have been brought to our attention in L2 research. A number of studies have reported that these constructs are among the most reliable predictors of L2 proficiency and L2 speech performance (De Jong et al., 2013; Iwashita et al., 2008; Révész et al., 2016; Zareva et al., 2005). In addition, within psycholinguistic and cognitive perspectives to second language acquisition, fluency and lexical complexity have been assumed to be inter-related in the speech production process. However, as will be explained later in this chapter, each of these is multi-faceted and a complex construct that is difficult to define or measure. Also, there have been major challenges in understanding how L2 learners develop their performance over time in regard to these aspects. To this end, L2 researchers aim to identify the factors that exert a potential impact on the development of fluency and lexical complexity in second language learning.

This chapter provides a review of the literature that has informed the theoretical underpinnings of Study 1 and Study 2; both studies are mainly centred on the examination of fluency behaviour in Turkish learners of English while the examination of lexical complexity behaviour forms only a part of Study 2. Therefore, allocating a larger space for the discussion of L2 fluency and important issues surrounding it, I present this review in three main sections. In the first section, I briefly explore L1 and L2 speech production models. Here, I aim to explain how L1 and L2 speech is produced and what differences exist between the two. Starting the section with a review of Levelt's (1989) model of L1 speech production, I then provide a brief discussion of Kormos (2006)'s bilingual speech production model. In the second section, I discuss aspects of speech performance and introduce the reader to a discussion of the two constructs: fluency and lexical complexity. For each of these, the components are outlined, and the main issues relating to the two constructs have been defined and measured in L2 research are presented. The aim in the second section is to help the reader understand why these two constructs are of particular importance in SLA research and how they play a pivotal role in understanding L2 speech and the relationship between L1 and L2 oral performance. The final section discusses the potential impact of three additional variables, i.e. individual learner variables of LP and WMC and LoR and an external factor of task structure, on the strength of the relationship between L1 and L2 fluency behaviours, if any exists. Finally, the chapter concludes with a brief summary of the key issues discussed throughout.

2.2. Speech production models

In studies investigating oral performance from a psycholinguistic perspective, the most widely cited and the best known theory of speech production is that of Levelt (1989), which is labelled as a 'blueprint of the native speaker' in a later version (Levelt, 1999), Levelt's model has since been revised and adapted by other researchers to explain L2 speech production (De Bot, 1982; Kormos, 2006; Segalowitz, 2010); therefore most L2 speech production models draw on Levelt's model for L1 speech production and share a conceptual structure based on this. In what follows, I first review Levelt's model and then an L2 speech production model, i.e. Kormos (2006). Drawing on the relevant literature, I also provide a discussion of the implications for Study 1 and Study 2 in each of the sections below.

2.2.1. L1 speech production

Currently the most widely accepted and empirically supported speech production model is Levelt's blueprint of the native speaker (1989, 1999) (Figure 2.1). Levelt created this model in order to understand and describe the spontaneous speech of a monolingual adult speaker. The model posits that in the speech production process, a speaker goes through four separate but interrelated stages: conceptualization, formulation, articulation and self-monitoring.

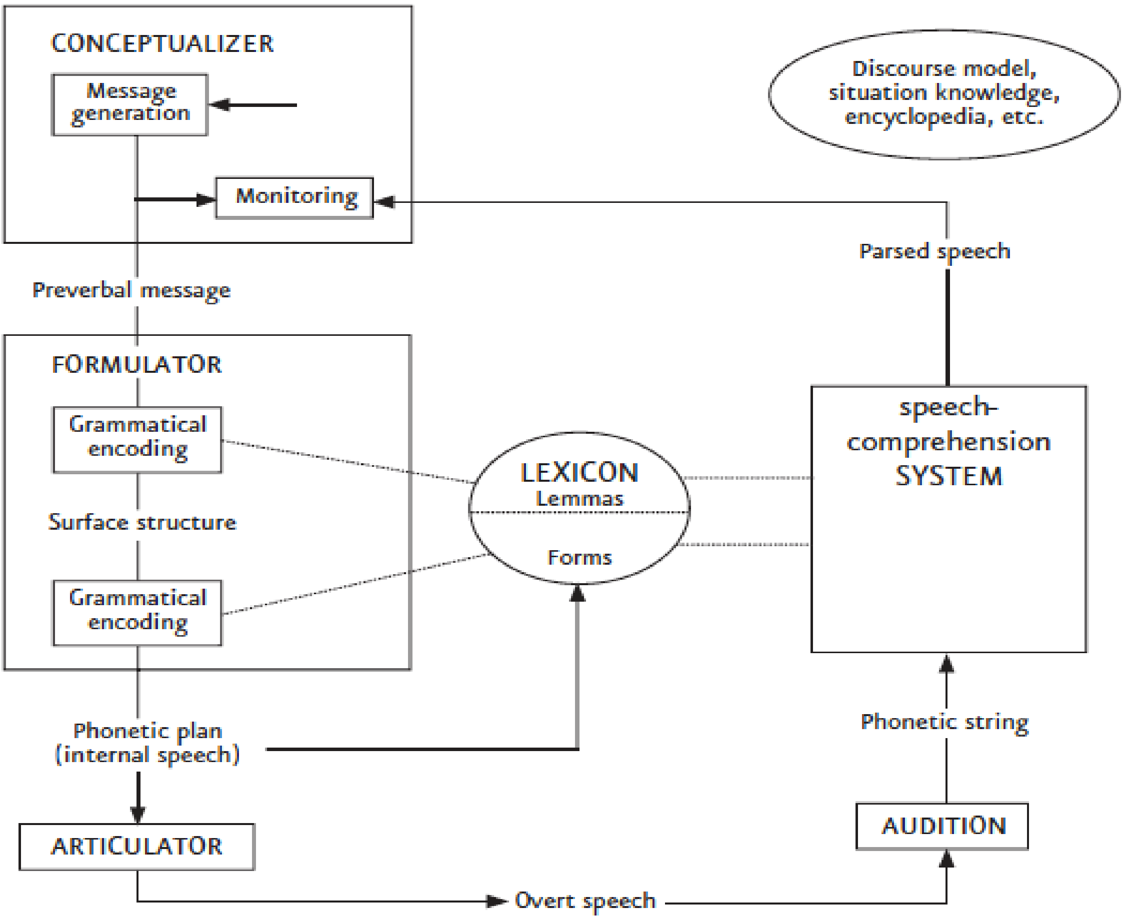


Figure 2.1. Levelt's (1999) Speech Production Model

The first stage, *conceptualization*, is where the speaker basically selects, organizes and sequences ideas. At this stage, the speaker engages in processes to decide what they intend to say (i.e. the communicative goal) and then develop the details of their intended message in two sub-processes: *macroplanning* and *microplanning*. Macroplanning involves the elaboration of the communicative goal (i.e. intended message) based on the speaker's knowledge of external and internal world, and making decisions about the register, i.e. informal or formal speech, and speech acts, such as requesting, apologizing, etc. This sub-process, however, is still not very detailed nor language specific. Microplanning, on the other hand, is more detailed and involves further shaping the message; i.e. the speaker considers the addressee and the message focus and decides on such points as the lexis to be used, perspective to be taken, old and new information, mood, etc. In fact, microplanning is where the intentions of the speaker are realized in the form of a so-called *pre-verbal message.* Pre-verbal message, the output of the conceptualization process, is not linguistic but still remains a conceptual structure. It contains all the necessary information about the message (e.g. the subject, the register and the lexis needed to produce the message satisfactorily) to turn it into a linguistic shape and can also be monitored. In micro-planning, the further conceptualization of the pre-verbal message requires one to organize and sequence ideas, which principally relates to discourse organization; thus, this sub-process could be assumed to be language specific.

In the second stage of the model, i.e. *formulation*, the preverbal message is taken as input and translated into a linguistic structure (i.e. phonetic plan) through the retrieval of appropriate grammatical, lexical and phonological structures according to the message content. The retrieval of such information involves grammatical encoding (i.e. by grammatical coder) and phonological encoding (i.e. by phonological encoder). The grammatical encoder is what gives the linguistic shape to the pre-verbal message; it accesses the lemmas, consisting of syntactic and semantic information, and the lexemes, consisting of morphological and phonological forms. These, i.e. lemmas and lexemes, are located in another knowledge source, i.e. *mental lexicon*, in the long-term memory (LTM), i.e. 'the store of information about words in one's language' (1989). The grammatical encoder selects these lexical items based on the speaker's intentions, and incorporates them into a surface structure, which is 'an ordered string of lemmas grouped in phrases and subphrases' (ibid., p. 11). Next, the phonological encoder comes into play, to convert the surface structure into a form that can be generated as overt (or vocalised) speech. In this way, a concept is formed into a propositional message, and this is represented as a phonological score in the system (Figure 2.1). So, the phonetic plan is the output of the formulation stage and provides the input for the next stage, i.e. *articulation*.

Articulation involves the execution of the phonetic plan by converting it into the final product, i.e. overt speech. For this process to take place, the plan (i.e. the internal speech) is first stored temporarily in the 'articulatory buffer', and then the articulator 'retrieves successive chunks of internal speech from this buffer and unfolds them for execution" (ibid., p. 13). The final product of the process, 'overt speech' (ibid., p.13), is finally produced through the articulatory organs, i.e. 'the lips, tongue, teeth, alveolar palate, velum, mouth cavity and breath' (Bygate, 2001).

A further stage in Levelt's model is *monitoring*. Levelt's postulates that all messages (i.e. pre-verbal and internal speech) are stored in the short-term memory (i.e. working memory), and that the speaker can monitor these while they can simultaneously monitor other people's overt speech. Monitoring, in three different loops in the model, helps the speaker detect errors before or after the execution of the overt speech; the speaker can check 1) the intended messages *during the conceptualization stage*, 2) the internal speech *during the formulation* before the speech is transferred to the articulation, and 3) the overt (or vocalized) speech, when the speaker hears it, to see if it is suitable for the purpose of the message.

Regarding these different but interrelated sub-processes in the model, two important points are brought into our attention. Firstly, Levelt's model is based on psycholinguistic research, and therefore, a distinction between automatic and controlled processing constitutes a fundamental discussion in it. Controlled and automatic processing coexist in the production process. Formulation and articulation are claimed to be highly automatic (without conscious attention) in L1 speech while conceptualization and monitoring processes involve controlled processing. Conceptualization requires several decisions to be made, as just mentioned, which demands attention from the speaker while monitoring is also a controlled process since the speaker consciously makes corrections in their speech. The second point relates to the fact that speech production is modular in that it consists of separate modules each of which does its work and transmits it to the next stage. For instance, when a pre-linguistic message has left the conceptualizer and is being processed in the formulator, another one previously processed can be articulated. In this sense, there is no need for the conceptualiser to wait for the whole chunk to go through the whole system to start a new part. This indicates that the speech is produced in a serial and parallel manner. It could therefore be inferred that automatic and parallel processing takes place in L1 speech production process, with all parts of the system simultaneously attending to different parts of the process speech.

Levelt's model is important in that it provides a summary of possible 'linguistics, psycholinguistic and cognitive issues underlying the act of speaking' (Segalowitz, 2010, p.8) However, one of the limitations of the model, as also acknowledged by Levelt (1999), is that it does not provide a developmental perspective but rather describes a single point in time (De Bot, 1992; Segalowitz, 2010). It is also important to remember that the model was built to describe speech processes that a 'unilingual speaker' goes through. On the other hand, we know that an L2 speaker has the linguistic knowledge of more than one language, although to different degrees, and therefore is likely to be faced with several issues in the speech production system that are different than those an L1 speaker faces. For example, questions remain as to when the speaker selects the language to be spoken or what knowledge stores they use in L2 speech. Since the blueprint does not accommodate such points that are specific to L2 use, other researchers (e.g. De Bot, 1992; Kormos, 2006; Segalowitz, 2010) took the main tenants of Levelt's model and revised it by pointing out where L2-related processes or issues can be located in the model. In what follows, I discuss L2 speech production and how it differs than L1.

2.2.2.L2 speech production

Generally speaking, it is agreed that similar processes take place in L1 and L2 speech production; however, a few adaptations must be made to Levelt's L1 model to accommodate specific aspects of L2 speech production. By drawing our attention to the two stages of Levelt's model, i.e. conceptualization and formulation, where differences in L2 speech production process lie. These models postulate that in conceptualization (more specifically macroplanning), the language to be used in speech is selected, and that differences also exist in the way that encodings take place in *formulation*. While encoding in L1 is automatic, making L1 speech smooth and fast, L2 speech requires conscious attention from the speaker due to the limited knowledge stores existing in L2.

A further difference between L1 and L2 speech production relates to how the linguistic knowledge is stored in the two; i.e. whether there is a separate store for L2 or shared store for L1 and L2 (Segalowitz, 2010). While there are several bilingual models in second language acquisition literature (De Bot, 1992; Segalowitz, 2010), below I review

Kormos's (2006) bilingual speech model and provide a detailed account of how L2 speech differs from L1 speech. I then discuss the main implications these have for Study 1 and Study 2.

2.2.2.1. Kormos's (2006) bilingual speech model

Revising Levelt's model, Kormos (2006) devised a bilingual speech model. According to her model, speech processing in L1 and L2 share the main stages (i.e. conceptualization, formulation, articulation and monitoring) and follow the same order but some variations exist in the way processing takes place. One important difference relates to the existing knowledge stores in each. Levelt's model has three knowledge stores, i.e. store of the knowledge of world, the mental lexicon, and a syllabary. The first of these includes information about internal and external world such as interlocutor or discourse conventions (e.g. choice of formal or informal register). Using the information here to produce something to say, the speaker begins the production process (in the conceptualization stage). The mental lexicon is located in the long-term memory and involves information about words of a language such as lemmas or morpho-phonological codes. This information is retrieved from the lexicon during the grammatical and morpho-phonological encodings in formulation. The syllabary, on the other hand, is "a repository of gestural scores for the frequently used syllables of the language" (Levelt, 1999) and is accessed in the phonetic encoding of the internal speech. In the case of L2, however, Kormos (2006) posits that there is an additional knowledge store, i.e. declarative knowledge, which contains L2-specific linguistic knowledge in learner's system, e.g. syntactic, phonological and grammatical rules of an L2. Kormos proposes that the information retrieval in L2 encoding process happens through this specific knowledge store. However, the rules here, which are typically stored in the mental lexicon in L1 speech and retrieved automatically, have not yet become automatized in L2 case.

Kormos claims that all the knowledge stores mentioned above are shared between L1 and L2, except the declarative knowledge store. Other researchers (Bolibaugh & Foster, 2013; Skehan, 2009, 2015; Skehan, Foster, & Shum, 2016) also agree with Kormos in that the main difference between L1 and L2 speech production lies in the nature of the mental lexicon. For example, Skehan (2015) explains that while L1 lexicon is "extensive, but equally important, well organized with rich entries often multiple storage of the same item", in the L2 case, it is much smaller and has a 'superficial nature' with lacks extensive links to other stores (p.127). It is understood that the declarative knowledge is rather limited and not yet fully automatic (at least for most learners), and this seems to be the main reason for the processing difficulties L2 speakers face in lemma retrieval and encoding processes. Encoding messages in the formulation (as well as articulation) stage requires more conscious attention and takes more time for speakers, especially for those at lower-levels of proficiency or when low-frequency words are retrieved. Therefore, when faced with such challenges, learners are assumed to engage in controlled processing depending upon how automatized their knowledge has become. This suggests that the degree of automaticity is an important factor here affecting the processing speed in bilingual speech production.

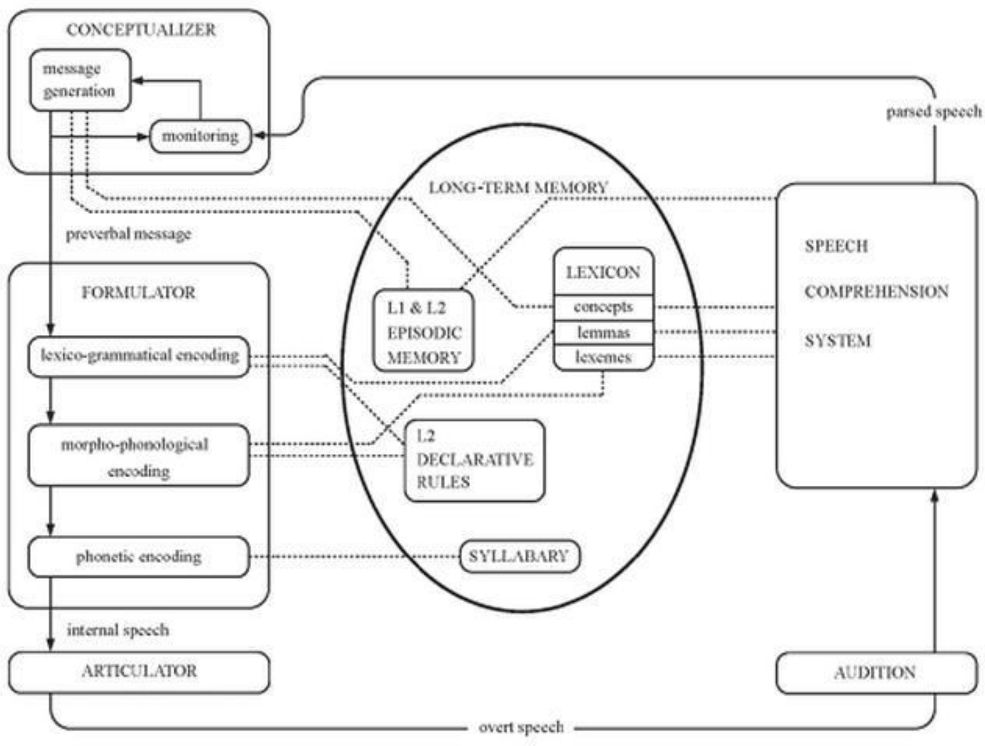


Figure 2.2. Kormos' bilingual speech model (2006), p.168

By automaticity it is meant that the declarative knowledge (i.e. linguistic information about a language learnt) becomes proceduralized through repeated practice; learners could apply it automatically without effort or paying conscious attention to it. L1 speech already relies largely on automatic and parallel processing; yet, L2 speakers must go through an effortful processing because they face additional problems in the flow of the process due to incomplete linguistic knowledge or low proficiency level in L12 acquisition. This also explains why L2 speech is not parallel but serial; unlike L1 speakers, L2 speakers cannot attend to all components of the system simultaneously, but instead have to attend to each separately. However, as Kormos adds, parallel processing may be possible for advanced learners as their declarative knowledge stores are more extensive and they develop higher degrees of automaticity in access to those stores.

2.3. Implications for Study 1 and Study 2

Having introduced a description of how processing in L1 and L2 speech are assumed to take place within a psycholinguistic view, I would now like to draw the reader's attention to two major implications that could be drawn from the above discussion for Study 1 and Study 2. First, the above discussion has suggested that while the way L2 speakers attend to conceptualization is similar to L1 speakers, it is the formulation that becomes more demanding and vulnerable for them. This is assumed to be due to two reasons: 1) L2 knowledge stores are not rich enough or the attained proficiency level is low, and 2) L2 speaker's access to these stores has not yet become fully automatic. So, the conceptualization and formulation stages could be considered separately in the L2 case since learners are assumed to engage in these differently.

The second implication relates to the inter-relatedness of fluency and lexis aspect of speech production. I have mentioned that the parallel functioning of L1 speech is essentially very effective; the demands of the preverbal message on the formulator (e.g. access and retrieval of the linguistic information from the mental lexicon or morpho-phonological encoding), can be met without difficulty in L1 speech. However, in L2 speech production, when the lexical demands (e.g., access and retrieval) cannot be met by the mental lexicon, processing problems occur, and the smoothness of the speech is disrupted. In other words, due to the less-developed nature of the lexicon in L2, particularly at lower levels of development, dysfluencies are likely to surface in the internal as well as overt speech. The overt speech becomes slower with frequent hesitations and pauses, especially within clauses (De Jong, 2016; Skehan, 2014; Tavakoli, 2011). Therefore, one main assumption frequently made is that lexis and fluency are interlinked in the sense that challenges in identifying, accessing and retrieving lexical items (whether low or high frequency) in the lexicon manifest themselves as occurrences of disfluencies in the overt speech.

Although these are the main assumptions underlying the examination of L2 oral performance in Study 1 and Study 2, at this point, I should make two important acknowledgements. First, formulation is not the only stage in the speech production system where L2 speakers encounter difficulties which lead to disfluency. For example, a topic shift in the conceptualization is also likely to disrupt the process leading to disfluency (see Segalowitz, 2010 for a detailed discussion of several points in the system which are vulnerable to disfluencies). Second, it cannot be asserted that the processing difficulties are not the result of one single factor such as a smaller L2 lexicon. Some other factors, for example, that should be considered include the degree of automaticity in encoding which is associated with L2 proficiency level (Kormos, 2006, 2011) or particular demands placed by the nature of specific tasks such as task design features (De Jong & Vercellotti, 2016).

I have mentioned that there are variations between L1 and L2 speech processing. One such variation could also be explained by individual differences as L1 personal speaking styles, WMC, the amount of L2 experience in the TL community (i.e. LoR), or by cross-linguistic differences between languages. Such factors should be also considered when examining L1 and L2 speech processing. As I discuss further in detail, such individual factors have been shown to be interacting with not only language acquisition processes in general but with speech planning and language production as well. In the following sections, I turn to some of these; I explain the factors that are within the scope of Study 1 and 2; namely, three individual learner variables, i.e. L2 proficiency, WMC and LoR, and one external factor, i.e. task structure. However, before moving on to a discussion of these, I first elaborate on the two aspects of speech performance, i.e. fluency and lexical complexity. These are the foci of this research because they are assumed to be reliable key indicators of L2 proficiency, which are interlinked to each other. My discussion also involves explaining why examining the relationship between L1 and L2 oral performance regarding these two aspects is important and how this contributes towards a better understanding of the differences in L1 and L2 speech processing systems.

2.4. Aspects of second language oral performance

In recent decades, a triad framework known as CAF (complexity, accuracy and fluency) has emerged in response to a need for a clear definition of what ‘speaking’ entails in research, assessment and teaching practices. Skehan (1996, 1998) was one of the pioneers in the field who brought CAF into task-based research and suggested that these performance dimensions could be measured separately. Since then, a large number of researchers have adopted this framework to examine successful second language oral performance (Housen & Kuiken, 2009; Towell, 2012) and have examined these aspects extensively in their research as a means towards describing and measuring development, proficiency or acquisition in L2 (De Jong et al., 2012b; Foster & Tavakoli, 2009; Révész et al., 2016; Tavakoli & Foster, 2011). Although lexis was already included as part of complexity in the framework, in a later study (Skehan, 2009), Skehan suggested that since lexis has a central role in Levelt’s model, it should be incorporated as a separate area into these performance descriptors. Therefore, what was originally known as CAF has become CALF (Complexity, Accuracy, Lexis and Fluency) and been referred to as such since.

The components of CALF, which are inter-related but can be measured separately, have been found to be theoretically and empirically valid measures that capture different aspects of L2 performance and proficiency (N. C. Ellis & Larsen-Freeman, 2006; Norris & Ortega, 2006, 2009; Skehan, 2003). The constructs have been widely examined in studies, as dependent variables, to explore the influence of a wide range of other variables (e.g. age, task variables or individual differences) on L2 performance (e.g. (Foster & Skehan, 1996; Freed, Segalowitz, & Dewey, 2004; Kuiken & Vedder, 2007; Robinson, 2011; Skehan & Foster, 1999; Yuan & R. Ellis, 2003) as well as independent variables to understand the psycholinguistic processes underlying L2 development and performance (e.g. Bosker et al., 2013; De Jong et al., 2012b; Hilton, 2008; O’Brien, Segalowitz, Collentine, & Freed, 2007; Skehan & Foster, 2007). CALF indices have also been adopted in important language frameworks (e.g. ACTEFL or CEFR) which formed the basis of benchmarks in high-stake international tests (e.g. IETLS, TOEFL) or informed materials in L2 language teaching.

Despite the popularity of these measures across L2 studies, several researchers expressed their concerns over the way these were employed across studies (Housen & Kuiken, 2009; Lambert & Kormos, 2014; Pallotti, 2009). Criticisms of CALF centre around three main issues. The first of these pertains to the interaction of the constructs with each other (R. Ellis, 2008; Housen & Kuiken, 2009; Skehan, 2008). It is rather difficult to claim that these are distinct from each other as they seem to be influenced by each other as well as by additional factors. For example, an increase in fluency development may be at the expense of accuracy development since these knowledge areas have differential developmental trajectories in the process of L2 acquisition (R. Ellis, 1994, 2008; Skehan, 1998). Or within cognitive perspectives, e.g. those that see humans as limited in attentional resources, these areas can compete for attentional sources; i.e. accuracy competes with complexity while fluency with accuracy (Skehan, 2015; Skehan & Foster, 1999).

Second, the way these areas of performance have been defined across different studies is problematic. Housen et al. (2012) point out that many researchers do not define meanings of these terms in sufficient detail, and maybe more importantly, existing definitions are not informed by theories of language learning or linguistics. This puts the validity of the constructs under question. Furthermore, there have been issues concerning the operationalization of CALF; researchers have measured these using various tools and ways ranging from identifying temporal features to holistic and subjective ratings by experts. Yet, some researchers have been critical about the reliability and validity of such measures and tools as often these are not straightforward (Norris & Ortega, 2008; Ortega, 2003). Equally important, due to the variability in definitions and operationalizations, studies have yielded inconsistent findings.

Despite the existing difficulties in defining CALF, I find it useful to provide Housen and Kuiken's (2009) definitions here. They suggest that complexity (of both syntactic and lexical) and accuracy are primarily related to 'the current state of the learner's (partly declarative, explicit and partly procedural, implicit) interlanguage knowledge (L2 rules and lexico-formulaic knowledge)' whereas fluency is seen to be primarily linked to 'the learner's control over their linguistic knowledge' (p. 2). It is understood that accuracy and complexity are largely related to the declarative knowledge; fluency, on the other hand, is linked to implementation of that knowledge and it manifests itself in how easy and fast the speaker accesses the relevant information in real time communication (ibid.). In this regard, fluency could be suggested to increase along with automaticity in speech production system (Wolfe-Quintero, Ingaki, & Kim, 1998).

Discussions of the next sections will centre on fluency and lexical complexity aspects only since these are the two constructs within the scopes of Study 1 and Study 2. Although I acknowledge that the examination of accuracy and syntactic complexity from a cross-linguistic perspective is valuable as they also contribute to our current understanding of L2 oral performance, only fluency (in Study 1 and Study 2) and lexical complexity (in Study 2) are the foci here for two reasons: 1) these two aspects have usually been reported as stronger indicators of LP in L2 studies as discussed in previous sections (De Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2012a; Gilabert & Munoz, 2010; Iwashita et al., 2008; Malvern & Richards, 2002; Révész et al., 2016; Yu, 2010; Zareva et al., 2005), and 2) the present research is built on the assumption that fluency and lexical complexity are possibly inter-linked in that the processing difficulties (e.g. lexical retrieval problems) encountered in formulation are likely to result in disfluencies in the actual performance. In what follows, I will first review and evaluate various definitions of fluency and this will lead up to the working definition of fluency in the current research. Then, a discussion of previous studies on L1 and L2 fluency and their implications for the present research will be provided.

2.4.1. Fluency

Fluency is a complex and multi-componential construct which interacts with and encompasses several different factors (e.g. linguistic, psycholinguistic and sociolinguistic) (Kormos, 2006; Lennon, 2000; Segalowitz, 2010). As such, a considerable amount of research has been devoted to examining what fluent speech entails (Derwing et al., 2009; Lennon, 2000; Segalowitz, 2010) or how it can be developed (Koponen & Riggenbach, 2000; Lennon, 1990). Fluency has been an important area of focus in L2 research for three major reasons: 1) it is one core area of L2 oral performance (i.e. CALF) (Housen & Kuiken, 2009; Housen et al., 2012; Skehan, 1998, 2003), 2) it has been closely related with L2 proficiency as one of the most reliable predictors of it (De Jong et al., 2012a; Révész et al., 2016) and 3) it has been seen a means towards an in-depth understanding psycholinguistic processes underlying L2 oral performance and development (e.g. gaining automaticity) (Kormos, 2006; Segalowitz, 2010). Therefore, exploring fluency is necessary to shine more light on our current understanding of L2 oral performance, L2 proficiency and the underlying cognitive processes involved in L2 speech. Yet, how do we define this concept? This is what I explore next.

2.4.1.1. Definitions of fluency

In L2 research, several definitions have been proposed to describe L2 fluency, which are in large based on L1 definitions and usually refer to the smoothness and fluidity of the speech. Simple though they might seem, the meanings of such terms as smoothness or fluidity remain rather vague and should be defined and operationalized clearly. Below I provide some of the existing fluency definitions. Though these are amongst the most accepted and widely cited ones, fluency definitions should by no means be restricted to the ones reviewed below.

2.4.1.1.1. Fillmore (1979)

One of the earliest studies on fluency was that of Fillmore (1979). Departing from how we judge L1 speakers, he outlined four abilities that speakers should have to be considered fluent: 1) the ability to speak at length with few pauses 2) the ability to talk clearly and concisely (i.e. coherently), 3) the ability to talk appropriately on a range of topics and 4) the ability to be creative with the language (i.e. using it creatively) (p. 93). Fillmore refers to L1 speakers when describing fluency. According to these abilities, a native speaker of language would be considered fluent if they produce speech which is long with few pauses, clear and coherent. They also need to be creative with language and demonstrate their knowledge and ability to speak about various things. However, this definition does not seem appropriate for an L2 speaker for two reasons. First, each of these abilities attempts to define fluency from different perspectives along the line of its multi-dimensional nature. It is not clear how each ability contributes towards defining the construct or which ones do more so. For example, could a speaker who speaks without hesitation but cannot speak on a variety of topics be considered fluent? Or is the ability to be creative with the language a more accurate or better representation of a fluent speaker than the ability to talk coherently? Fillmore's definition does not provide answers to such questions.

More importantly, the abilities described above are rather vague and open to different interpretations. To give an example, coherence could be related to the content of the speech, which should be logical, or it could be related to the content and articulation, i.e. how the speech is produced by the interlocutor. What is more, the nature of the speech is highly dependent upon the context; e.g. a speech at a steady speed without pauses could be unlikely or undesirable within group discussions as pausing a lot in such contexts would be perfectly natural. Similar criticisms over Fillmore's definition of a fluent speaker are also echoed by other researchers (Thomson, 2014; Witton-Davies, 2014).

2.4.1.1.2. Lennon (1990, 2000)

In contrast with Fillmore, Lennon (1990) takes an L2 perspective to defining fluency. Lennon defines fluency as 'an impression on the listener's part that the psycholinguistic processes of speech planning and speech production are functioning easily and efficiently' (p. 391). Here, Lennon refers to what the listener can infer, judging from the observable performance, about the speaker's cognitive processes underlying his speech.

However, Lennon does not directly or explicitly address the potential link between cognitive processes and speech fluency, but rather he aims to identify how fluency is manifested in the overt speech.

Also, Lennon makes a distinction between two senses of fluency: a broad sense and a narrow sense. In the broad sense, fluency is taken to be ‘a cover term for oral proficiency, and it represents the highest point on a scale that measures spoken command of a second language’ (p. 389). In this view, fluency seems to be an equivalent of global oral proficiency and covers other aspects of performance (e.g. lexical variety, syntactic complexity or accuracy). In the narrower sense, on the other hand, fluency refers to one component of *oral proficiency*, often used as scores in oral performance assessments. This aspect usually refers to the temporal aspects of language (e.g. the amount of language produced per unit of time) and is commonly used in L2 research contexts to describe non-native speakers’ speech fluency. As Lennon (1990) puts it, fluent speech in the narrow sense is defined as speech which is ‘unimpeded by silent pauses and hesitations, filled pauses (‘ers’ and ‘erms’), self-corrections, repetitions, false stars and the like’ (p. 390).

In a later study, Lennon (2000) proposed one of the most comprehensive definitions of fluency, which is ‘the rapid, smooth, accurate, lucid and efficient translation of thought or communicative intention into language under temporal constraints of online processing’ (p. 26). As was in his previous definition, in Lennon (2000) fluency seems to be a performance phenomenon with a particular focus on temporal aspects (e.g. pauses, speech rate). The terms are, however, still somewhat vague; for example, it is not clear what ‘lucid, accurate or smooth’ specifically mean in this context. One could argue that adjectives/concepts such as lucid and smooth are relative terms and standards for lucidity and smoothness may vary among different listeners. Therefore, the definitions of such terms need to be defined clearly.

In Lennon’s definition, the vagueness is compounded by two further issues: 1) it entails ‘accurate’ language and 2) it implies ‘automatic processing (i.e. by ‘the efficient translation of thought into language) and no conscious planning (by ‘under the temporal constraints of online processing’). Both of these points could be considered problematic.

Regarding the first issue, in the CALF framework, accuracy is treated as a distinct area of performance and thereby is examined on its own (Housen & Kuiken, 2009; Housen et al., 2012). Second, fluency is seen as speaking effortlessly without conscious planning. Yet, high proficiency learners could still produce fluent speech but may rely on conscious planning and processing (Kormos, 2006). Thus, the nature of Lennon's revised definition could still not be considered as unproblematic.

2.4.1.1.3. Segalowitz (2010)

Segalowitz (2010) recognizes that there is ambiguity and variability in the way fluency is defined and operationalized, and that this variability brings challenges first in conceptualizing the construct and second, in its measurement. In order to identify the reliable indicators of fluency in L2 oral performance, Segalowitz approaches fluency from a cognitive perspective and asks the following question:

> 'What features of L2 oral performance serve as reliable indicators of how efficiently the speaker is able to mobilize and temporally integrate , in a nearly simultaneous way, the underlying processes of planning and assembling an utterance in order to perform a communicatively acceptable speech act?' (Segalowitz, 2010, p. 48)

Inasmuch as a question it is, the above quote also entails how Segalowitz views fluency. He sees fluency as the extent a speaker can efficiently mobilize and temporarily integrate underlying speech production to perform an acceptable speaking task. He focuses on the links between cognitive processes which are functioning in speech production and the certain characteristics of the speech produced/articulated. That is to say, he shifts the focus from the observable characteristics of '*fluent speech'*, to '*fluency'* itself (p. 48). Consequently, he proposes a conceptualization of fluency in a triadic framework, in which each fluency type is distinct but interrelated to others: *cognitive fluency*, *utterance fluency* and *perceived fluency*.

Segalowitz defines *cognitive fluency* as "ability to efficiently mobilize and integrate the underlying cognitive processes responsible for producing utterances" (Segalowitz, 2010, p. 48). He suggests that speech involves several different interacting cognitive processes, and these processes should be coordinated efficiently in the system in order to ensure that the intended utterance is produced quickly and without disrupting the flow of the speech. The coordination of such activities includes mechanisms for speech planning, for lexical search and/or access, and for producing an internal speech for the articulation of the speech. To ensure cognitive fluency, these sub-processes in the speech production system should take place with minimal interference from internal sources or 'crosstalk' that could affect the processing (p. 48).

While 'cognitive fluency is the fluency that a speaker possesses', *utterance fluency* refers to 'fluency characteristics that a speech sample can possess' (p. 48). Certain features of speech, i.e. the temporal characteristics such as speed, pauses or hesitations, can be observed in the performance and measured objectively. In this sense, fluency is not related with the impression hearers might have about the utterances. Segalowitz intends to interpret utterance fluency in a much narrower sense and refers only to those features in speech which are directly linked to cognitive fluency, though he also acknowledges that no evidence of such links have been presented yet. The definition of utterance fluency here could actually be suggested to coincide with narrow definitions of fluency, i.e. one aspect of speaking ability and temporal features of a speech sample. This is also implied in how he defines utterance fluency as 'fluidity' and maintains: 'This fluidity is a property of the actual speech and is reflected in its temporal characteristics (speech rate, patterns of hesitations, etc.)' (p. 163). Regarding what features of speech exactly might be considered as important for utterance fluency, Segalowitz further notes that theoretically there might be a wide range of ways that features of a speech sample could be operationalized; the example of this could be speech rate (i.e. it can be operationalized as syllables per second or between pauses). Although this issue (i.e. there being a wide range of measures used in L2 research to operationalize observable features of speech) will be discussed further in the next sections, for the moment, what needs to be pointed out is that utterance fluency, according to Segalowitz (2010), is distinct from cognitive fluency, and therefore they should be examined separately.

The third category of fluency *is perceived fluency*, which concerns the impression on the part of the listener about the cognitive fluency of a speaker. In other words, listeners have their own perceptions of utterance fluency, and based on these, they make inferences about speakers' cognitive fluency. In this sense, listeners might not typically identify every pause or hesitation in what they hear, nor might they treat these as instances of disfluency. This fact actually might imply that hesitations or pauses are normal and even likely, to some extent, in fluent speech. Given that most definitions of fluency refer to the absence of undue hesitations or pauses, it seems important and necessary to distinguish this type of fluency from others (i.e. cognitive fluency and utterance fluency).

Once again, these three types of fluency are theoretically distinct but not independent of each other; for instance, cognitive fluency is the underlying processes that affect utterance fluency, and utterance fluency in turn affects listeners' perceptions of fluency. When L2 speakers have less developed automaticity in their speech processing (i.e. indicative of cognitive fluency) and have to speak in real time or with time pressure, they may find it challenging to speak fluently or to cope with the demands of the interaction. L2 speakers may or may not be aware of their inefficient processing in such cases but they need to stall for time, by employing a number of problem-solving or 'coping mechanisms (e.g. communication strategies, meaning-negotiation mechanisms, hesitation devices, repair mechanisms)' (Dörnyei & Kormos, 1998). For example, they can leave the message incomplete or reconstruct it by simplifying it, or some parts of the message can be re-coded in the speech production system (Kormos, 2006; Segalowitz, 2010). Entering a prolonged discussion of such mechanisms is not possible nor necessary here due to space and scope limitations. It is worth mentioning though that such problem-solving mechanisms related to time-pressure form the basis of reliable indicators of utterance fluency. These involve non-lexicalized pauses (i.e. unfilled pauses and umming and erring such as err, uh, uhm, and sound-lengthening), lexicalized pauses (i.e. fillers such as well, you know, okay) and repetitions (i.e. self-repetition or repeating what the interlocutor said) (see Dornyei and Kormos, 1998, p. 370, for a detailed discussion and examples).

The above discussion suggests that the distinction between three types of fluency offers a more coherent and systematic approach to defining and conceptualizing fluency. It also clearly indicates that the fluency construct is complex and multidimensional. I should note that following from Segalowitz's (2010) conceptualization of fluency, although still considered limited by some researchers (Kahng, 2014), Study 1 and Study 2 is interested in examining the utterance fluency only, i.e. observable features of the speech that can be measured objectively although Study 1 is also partly interested in cognitive fluency through the examination of task structure effect. In these studies, fluency is limited to its narrow definition, i.e. one aspect of speech ability amongst several others (e.g. accuracy, lexis, grammar, etc.) rather than general proficiency (Lennon, 1990) because 1) it is the only aspect that can be the most readily and objectively measured and 2) it is in this narrower sense that fluency has widely been investigated in L2 studies to describe non-native speakers' speech (Bosker et al., 2013; De Jong, 2016; De Jong et al., 2015; De Jong et al., 2012b, 2013; Derwing et al., 2009; Kormos & Denes, 2004; Tavakoli et al., 2019). Further review will be directed to utterance fluency with links to the other two aspects where relevant.

2.4.1.2. Utterance fluency

As already discussed, one way to differentiate fluent and less fluent speakers is to examine utterance fluency which shows features of speech indicative of cognitive processes underlying speech production. Utterance fluency (which will be referred simply as fluency henceforth) is typically believed to comprise at least three aspects: breakdown, speed and repair (Skehan, 1998, 2003; Tavakoli & Skehan, 2005). *Breakdown fluency* relates to pausing behaviour of speakers and is usually examined by measuring frequency and length of pauses, which could be filled and/or unfilled, per unit of time (De Jong, 2016; Derwing et al., 2009; Derwing, Rossiter, Munro, & Thomson, 2004; Rossiter, 2009; Tavakoli, 2011) while *speed fluency* is related to the rate with which the speech is delivered (e.g. number of syllables per second) (Kormos, 2006; Kormos & Denes, 2004; Peltonen, 2018). *Repair fluency,* on the other hand, is reflected through speakers' tendency to repair themselves during online communication (e.g. by rephrasing or correcting themselves). Instances such as repetitions, replacements, reformulations and false starts are assumed to be revealing repair processes underlying speech production (Dörnyei & Kormos, 1998; Hunter, 2017; Kormos, 1998, 2006; Tavakoli et al., 2019).

So, one important question that could be asked here is which measures are reliable indicators of fluency for each of these aspects. It has been a challenging task for L2 researchers to determine which features to look at in speech to characterize fluency, and they have taken several approaches to this end. Some studies have looked at L2 speech samples before and after an intervention on fluency and compared the two (Lennon, 1990; Towell et al., 1996) while some others have compared L2 speakers' speech with those of L1 speakers (i.e. using native speaker data as a benchmark) (De Jong, 2016; Skehan et al., 2016; Tavakoli, 2011). The latter approach was usually based on the assumption that differences between L1 and L2 speech data could imply processing differences underlying L1 and L2 speech production. Another way has been to correlate some speech features with the ratings of native speakers of the TL or non-native speakers, who are proficient users of the TL such as language professionals or practitioners (Derwing et al., 2004; Freed, 1995; Kormos & Denes, 2004). In such studies, native-speakers and/or non-native expert users are asked to judge L2 speech samples with or without a prior training on fluency, and the focus has generally been on examining perceived fluency. Now I will review how breakdown, repair and speed aspects have usually been operationalized and measured across studies.

2.4.1.2.1. Breakdown fluency

Breakdown fluency is associated with occurrences of pauses in speech. Yet, should we examine every single pause in speech to measure breakdown fluency? Earlier studies examined pauses which were 0.1 second (Foster & Skehan, 1996) or 0.40 second (Freed, 2000) while more recent studies suggest that pauses longer than 0.25 sec are better indicators of pausing behaviour as this length is considered to be a noticeable silence length for native-speakers of English (De Jong, 2016; De Jong & Bosker, 2013; De Jong et al., 2015; De Jong et al., 2012b; Huensch & Tracy-Ventura, 2016; Peltonen, 2018).

Pauses could be explored in terms of frequency, length, location (i.e. mid-clause or end-clause) and quality (i.e. filled or silent). Regarding the location of pauses, several recent studies have provided evidence that what differentiated L2 speakers from L1 speakers seems to be where they paused (i.e. within/ between a clause) rather than how often they pause (De Jong, 2016; Tavakoli, 2011). It is suggested that L2 speakers tend to pause more often within clauses; in contrast L1 speakers tend to do so between clauses. This distinction, i.e. pausing within/between clauses, is also assumed to be amongst the important factors influencing perceptions of listeners. L1 and L2 speakers' tendency to pause at different locations could be attributed to the processing differences in L1 and L2 speech production systems. L1 speakers are likely to pause at the end of pauses because they need to conceptualize what they are going to say next, whereas L2 speakers are faced with processing difficulties, e.g. looking for the right words, due to their incomplete L2 linguistic knowledge and this might lead them to pause within a clause (Kormos, 2006; Skehan et al., 2016). Bringing a similar but slightly different perspective, some other researchers point at evidence suggesting that what makes L1 and L2 speech different is what words (i.e. low or high frequency) come before pauses (e.g. Bosker, Quene, Sanders & De Jong, 2014; De Jong, 2016) as pauses in L1 speech usually precede low-frequency words. Further related to this, Skehan et al. (2016) also draw a distinction between within and between clauses. They argue that disfluencies *within clauses* stem from problems occurring in the formulation stage and are evidence of micro-planning issues while disfluencies at clause boundaries are linked to discourse decisions taking place in the conceptualization stage and are associated with macro-planning issues. They characterise the former as *clause-level disfluency* and the latter as *discourse level disfluency*.

A final point of discussion about pause-related fluency concerns the quality of pauses (i.e. filled or silent). Both silent and filled pauses are assumed to be indicative of language processing demands, e.g. retrieval of low-frequency words (Ginther et al., 2010; Iwashita et al., 2008; Kahng, 2014, 2020; Schmitt, 2010; Segalowitz, French, & Guay, 2017) while filled pauses might also be used by speakers to facilitate communication (e.g. as a communicative strategy) (Dornyei & Kormos, 1998; Tavakoli et al., 2019).

2.4.1.2.2. Speed fluency

As for the speed aspect of fluency, there is robust research evidence that the rate at which speech is delivered is a reliable marker of fluency (e.g. De Jong et al., 2015; Kahng, 2014, 2020; Kormos & Denes, 2004; Révész et al., 2016; Segalowitz et al., 2017; Tavakoli et al., 2019). Measures of speed fluency (e.g. speech rate) also predicts perceived fluency, e.g. fluency ratings of native-speakers or advanced level non-native speakers (Kormos & Denes, 2004). In recent years, researchers have argued that breakdown and speed aspects of fluency have been confounded in studies, meaning that some features of speech could be indicative of more than one aspect of fluency. This suggests that some fluency measures overlap with others and therefore should be classified as composite measures (Hilton, 2014). An example of a composite measure is speech rate, i.e. the number of syllables produced in a given time. It is suggested that although this measure has most commonly been employed as a measure of speed fluency in fluency studies, it is not independent of pausing behaviour and thus could not be considered a pure measure of speed fluency (Hunter, 2017; Tavakoli et al., 2019).

2.4.1.2.3. Repair fluency

Indicators of repair fluency traditionally included instances of repetitions, replacements, reformulations, self-corrections, false starts or hesitations. It has been suggested that these measures are useful indices which reveal repair strategies that speakers employ in the speech production process in order to monitor and modify their utterances (Hunter, 2017; Kahng, 2020; Kormos, 1998; Skehan, 2003; 2009; 2015). These instances (e.g. corrections, repetitions, false starts or reformulations) could stem from the demands of different stages of the production system; e.g. due to the amount of information in conceptualization or lexical retrieval or encoding demands in formulation. Here, it is important to mention that some researchers claim that repair indices should be seen and calculated as measures of accuracy rather than fluency (Gilabert, 2005, 2007). Researchers subscribing to this view suggest that such features in speech (e.g. self-corrections, reformulations, replacements) indicate speakers' attempts at being accurate and improving their own speech. Although this perspective could be another useful approach to understanding repair behaviour, in order to be able to compare the findings with those reported in previous fluency studies, repairs are seen as measures of fluency in the present study.

Several concerns have been expressed by researchers (e.g. Hunter, 2017; Tavakoli et al., 2019) about the use of these measures to gauge repair fluency for two reasons. First, it is suggested that some of these repair measures are interdependent on each other; for instance, that a false start in performance is mostly likely to lead to a reformulation. Second, whether repair fluency measures can consistently distinguish performance at different levels of proficiency has come under question. A recent study by Tavakoli et al. (2019) examined all fluency aspects in speech data of 32 candidates, and they employed a comprehensive range of fluency measures, across different levels of L2 proficiency in order to see how each fluency measure performed in terms of distinguishing between L2 proficiency levels. The authors reported that while speed and some breakdown measures distinguished fluency between lower-levels (i.e. A2 and B1) and higher-levels (i.e. B2 and C1), the results for repair measures did not indicate a linear relationship between fluency and levels of proficiency, suggesting the interplay between fluency and proficiency might be more complex. They reported that repair behaviours began to develop when the participants reached B1 level, and it began to decrease once they had achieved an advanced level (i.e. C1). Therefore, as the authors interpreted the findings, when it comes to repair aspect of fluency, the picture appears less clear. One could also argue that whilst repair measures are valuable in that they still tell us about speakers' repair processes, it might be a better approach to examine total repair, combining all individual measures, to gauge speakers' repair behaviour.

Returning to the question of which measures are reliable indicators of fluency, Witton-Davies (2014) list speech features that emerged as most reliable in his study: number of pauses, length of pauses, location of pauses, mean length of run, speech rate, articulation rate, phonation time ratio and a selection of repair measures. Unfortunately, however, the lack of consistency persisting in L2 studies about the operationalization of each fluency aspect and even the way each is defined remains an unsolved issue (Tavakoli, 2016). In addition, the range of measures employed to represent fluency as a whole construct or different aspects of fluency varies considerably, presenting further challenges for researchers in making cross-study comparisons. As already implied, other unclear points relate to whether simply counting the markers of fluency (say, pauses or hesitations) could discriminate between L1 and L2 speech or whether all aspects of fluency could be considered as indicators and/or discriminatory of L2 proficiency. Given that fluency is influenced by various other factors (e.g. individual differences, features of speech elicitation tools), it is difficult to determine to what extent such fluency features contribute towards our current understanding of cognitive processes involved in L2 speech production. Surely, more fluency research examining the role of such additional factors is needed.

To this end, Segalowitz (2010), among many other researchers, proposes that an analysis of L1 speech samples (i.e. samples elicited from the same participants and in their L1) and contrasting it with their L2 speech samples could be one approach. In this way, disfluencies that are specific to L2 data and result from the processing difficulties encountered in the L2 speech production system can be distinguished from those that are related to L1 background or personal characteristics/styles. Examining the implications of L1 speaking styles for L2 speech fluency would not only contribute to reliability of L2 fluency measures but would inform L2 fluency practices (e.g. teaching and assessment) as well. More importantly, comparing L1 and L2 fluency would provide us with an opportunity to gain a more in-depth understanding of the similarities and differences between L1 and L2 processing. This would certainly help with the development of a better-informed model of L2 speech production. Although in recent years, researchers have shown an increasing interest in exploring L1-L2 fluency links, efforts in this regard have remained rather limited. Cross-linguistic studies comparing L1 and L2 speech data are scarce with only a few exceptions (e.g. De Jong et al., 2015; Huensch & Tracy-Ventura, 2016; Riazantseva, 2001). In what follows, I will review some of these since these are the most relevant ones to the present study.

2.4.1.3. The influence of L1 fluency on L2 fluency

L2 research has recently begun to recognize the importance of L1 influence on L2 fluency. Derwing (2017) suggests that the impact of L1 might be observed in several different ways. One of these concerns personal speaking styles of people, e.g. their preference of speech rate, or their tendency to pause within speech as also pointed out by other researchers (De Jong et al., 2015; De Jong & Mora, 2019). This implies that if people have a tendency to speak fast in their first language, they are more likely to be a fast speaker in their L2 as well. Similarly, if one tends to pause a lot in their L1, they are expected to pause frequently in their L2. In addition, Derwing (2017) notes that L1 influence on L2 fluency behaviour can also be reflected through cross-linguistic (e.g. morphological, syntactical) similarities or differences between L1 and L2. From a typological perspective, some language pairs are more similar than other language pairs; learners whose L1 is structurally similar to L2 might find it easier to improve their L2 fluency than those who have very distant L1. These examples raise the interesting questions of to what extent L2 fluency could be judged without taking the speakers' L1 background into account (be of their personal styles or cross-linguistic influences), or to what extent fluency is a characteristic specific to an individual speaker (i.e. an underlying trait of a speaker) or to a certain language (i.e. language-specific state) (Derwing et al., 2009; Derwing, 2017). An additional question that would naturally follow is how much of L2 fluency behaviour could be attributed to L1 fluency behaviour. Clearly, exploring the relationship between L1 and L2 fluency behaviour is essential in getting answers for such questions. Investigating L1-L2 links would also provide further evidence for the development of better-informed L2 speech production models, and would have valuable implications for L2 practices (e.g. research, teaching and testing) in that for example, language assessment practices could be required to refine their fluency assessment criteria by correcting L2 fluency measures for L1 speaking style (De Jong et al., 2015; Ginther et al., 2010; Huensch & Tracy-Ventura, 2016).

As already discussed, however, there are relatively few studies examining the possible relationship between L1 and L2 fluency behaviours (Towell & Dewaele, 2005; Towell et al., 1996), and the results from this line of research are far from being conclusive or generalizable. The limited body of cross-linguistic studies examining L1-L2 fluency links adopted different methodological approaches, including 1) comparing group differences between L1 and L2 fluency performance (e.g. Towell et al., 1996) 2), analysing the relationships between L1 and L2 measures by correlational analysis (e.g. De Jong et al., 2015; Huensch & Tracy-Ventura, 2016), 3) exploring the predictive power of L1 fluency measures over L2 fluency measures (e.g. De Jong et al., 2015), and 4) investigating the predictability of L2 proficiency from corrected measures of L2 fluency by partialling out

L1 effects (De Jong et al., 2015; Towell & Dewaele, 2005). I will now turn to a detailed discussion of the studies that can inform the development of Study 1 and Study 2 reported in this thesis.

Towell et al. (1996)

One of the earliest studies to investigate the relationship between L1 and L2 fluency is Towell et al. (1996). The authors examined fluency development in 12 English advanced learners of French by comparing participants' L2 fluency performance before and after residing abroad for 6 months, and by also comparing learners' performance in their L1 and L2. They examined five temporal variables (i.e. fluency measures); namely, speaking rate (i.e. the number of syllables produced divided by total speech time including pauses), phonation-time ratio (i.e. the percentage ratio of time spent speaking to the time spent to produce the entire speech), articulation rate (i.e. the number of syllables produced divided by total speech time excluding pauses), mean length of run (i.e. mean number of syllables produced between silent pauses of 28 seconds or above) and mean length of silent pauses. The results showed that learners improved in their speed-related fluency after residing abroad; gains were observed in learners' speech rate, articulation rate and mean length of run, amongst which mean length of run contributed most to the fluency development and to the increases in speech rate. However, the authors reported that when compared with L1 fluency behaviour, L2 learners' performance 'still lagged behind' L1 speech with the exception of two measures, i.e. phonation-time ratio and mean length of pauses. In fact, regarding phonation-time ratio, learners approximated their performance to their L1. An interesting finding in the study concerns the speech rate, with strong positive correlations reported between L1 and L2 speech rate as measured both before and after residing abroad (r = .97 and r = .73, respectively). The authors added that learners who had the highest speech rate before residing abroad tended to improve least in this aspect after residing abroad, a point explained as that L2 learners reach a plateau for their speech rate at a certain point in their L2 proficiency.

Rizantsieva (2001)

Another early study is by Riazantseva (2001) who explored language-specific and universal aspects of pausing behaviour (in terms of duration, frequency and distribution of pauses) in L1 Russian, L1 English and L2 English. The study involved an experimental group of 30 Russian speakers, who were learners of English at intermediate and advanced levels and a control group of 20 native speakers of English. The data were elicited through two tasks: a topic narrative, which was described as less structured in that it allowed the participants to choose the content as well as the means to express themselves, and a cartoon description task, which was described as a highly structured task as it offered minimal freedom to choose syntactical, lexical or semantic structures. Examining silent pauses of 100ms and above, the author compared differences in the pausing patterns between L1 Russian and L1 English, L1 Russian and L2 English, L1 English and L2 English. The results showed that L1 speakers of Russian and English had different pausing patterns in terms of the duration of their pauses; L1 Russian speakers made longer pauses than L1 English speakers. On the other hand, pause frequency and distribution for these L1 groups were similar with both groups making a similar number of pauses in general and similar number of pauses within constituents (i.e. clauses). It was concluded that lengths of pauses might be specific to a language while pause distribution and frequency might be a universal feature.

Riazantseva (2001) also compared L2 English speakers' pausing patterns at intermediate and advanced levels with those of L1 English speakers, and reported that advanced level the participants produced similar lengths of pauses as those of L1 speakers while pauses of intermediate level speakers were longer than those of L1 speakers. However, speakers at both levels produced significantly fewer pauses in general than L1 English speakers, and they produced more within-constituent pauses regardless of their proficiency levels. Riazantseva interpreted these findings as that speakers at advanced level managed to adjust their speech to their L2 English in terms of the lengths of pauses, but not in terms of the number and distribution of pauses.

Derwing et al. (2009)

In a longitudinal study, Derwing et al. (2009) investigated fluency ratings in L1 and L2 for 16 Mandarin and 16 Slavic language (i.e. Russian and Ukrainian) speakers, who were immigrant learners in Canada and were at a beginner level of L2 proficiency. The data were collected over a two-year period at three points: 2-months 10-months and 1-year after residing abroad. The participants narrated a picture story once in their L1 and three times in their L2, and their speech samples were rated on a nine-point scale (1= extremely fluent and 9= extremely disfluent) by trained judges: Mandarin, Russian and English native speakers (8 of each). While Mandarin and Russian speakers were asked to rate on participants' L1 fluency and English native speakers were always asked to rate on participants' L2 speech produced, focusing on the temporal factors (i.e. number of pauses, self-repetitions, false-starts). Correlations between the fluency ratings in L1 and L2 demonstrated significant relationships for both L1 groups, with a stronger correlation for Slavic group though, at 2-month time only. No significant relationship was observed at other data collection times (i.e. 10-month and 1-year). The authors further compared temporal measures (i.e. number of pauses per second, speech rate and pruned syllables per second) in L1 and L2 speech samples and observed that all measures correlated significantly with each other at the initial stages of L2 exposure for both L1 groups, but for Slavic group only at later stages. They concluded that the group differences might have been caused by several factors, including varying amounts of exposure to English or to a 'closer relationship between Slavic languages and English than between Mandarin and English' (p. 534).

In Derwing et al. (2009), Mandarin group made fewer gains in their fluency development over time than Slavic group. In fact, in a follow-up study with the same participants (Derwing & Munro, 2013), no or little improvement of fluency was observed for the Mandarin group even after 7 years of residing abroad whereas Slavic learners were perceived considerably more fluent from the point of 2-year to that of 7-year. Interpreting the findings of the two studies, Derwing (2017) suggests that it is plausible to assume that 'L1 was a factor here in that English and the Slavic languages (Russian and Ukrainian) are both Indo-European whereas Mandarin is unrelated to English' (p. 251).

De Jong et al. (2015)

Perhaps the strongest evidence about the link between L1 and L2 fluency behaviour comes from some recent studies (e.g. De Jong et al., 2015; De Jong & Mora, 2019; Huensch & Tracy-Ventura, 2016; Peltonen, 2018). In a cross-linguistic study, De Jong et al. (2015) employed two types of fluency measures (i.e. uncorrected and corrected measures for L1 behaviour) to examine the speech of the same groups of participants in their L1 and L2. The authors argued that L2 measures commonly used in L2 research may not be measuring L2 specific skills but could be testing personal speaking styles as well. As such, their aim was to investigate whether L2 measures, when corrected for L1 fluency behaviour by using residuals, could more accurately predict L2 proficiency than uncorrected measures. In order to understand cross linguistic differences and to be able to generalize the findings to different language groups, they investigated two typologically different languages as L1s, namely English and Turkish, and Dutch as L2. Eliciting speech samples thorough eight different tasks from intermediate to advanced learners, they used several measures to represent each aspect of fluency, i.e. *mean syllable duration*, *mean silent pause duration* (within and between AS-units), *repetitions* and *corrections*. The participants' scores on a vocabulary test were used as an estimate of their L2 proficiency.

De Jong et al., (2015) reported that all L2 measures were, to a certain extent, related with L1 fluency, with the amount of variance ranging from 21 % to 57%. The results also indicated that all measures (both corrected and uncorrected) predicted L2 proficiency significantly, except the corrected measure of *mean syllable duration*, which predicted L2 proficiency more strongly than its original uncorrected equivalence. Also, *mean syllable duration* was found to be different across the two L1s, which the authors attributed to the typological differences between the two languages (i.e. English and Turkish). In other words, they explained the difference by the fact that syllables in Turkish tend to be shorter and thus the mean syllable duration is likely to be shorter in Turkish than in English. Another difference observed between the L1s was *the number of silent pauses*; with L1 Turkish speakers pausing less than L1 English speakers. The authors also explain this by typological differences between Turkish and English; since Turkish is an agglutinative language, where words can turn into longer words with the chains of morphemes, it is possible that Turkish speakers have fewer opportunities to pause.

De Jong and Mora (2019)

In a similar vein, De Jong and Mora (2019) explained that a substantial part of individual differences that surface similarly both in L1 and L2 fluency behaviour in semi-spontaneous speech could be related to personal ways of speaking which are language independent. The authors argued that speech production process involves the use of skills in each sub-process (e.g. conceptualization, formulation, articulation and monitoring) and individual differences in both L1 and L2 speech originate from how well these skills are executed in each sub-process. They were particularly interested in the skills used in the articulation stage, so they examined the relationship between articulatory skills (i.e. speech motor skills seen as language-independent) and individual differences coming forth in L1 and L2 speech fluency. In addition to this, De Jong and Mora (2019) aimed to examine the relationship between L1 (Spanish) and L2 (English) fluency behaviour. Working with 51 participants, who were at upper-intermediate level measured through a vocabulary size test, they examined the participants' articulatory skills and fluency features in their speech. Six picture-based tasks were utilised in L1 and L2 to investigate fluency and two other tasks (i.e. delayed picture-naming task and speeded syllable production task) to examine articulatory skills.

The authors reported strong correlations between L1 and L2 fluency for mean syllable duration, silent pause rate and mean silent pause duration. Their results suggested that articulatory skills could explain only a small portion of the variability in L1 and L2 fluency (only for silent pause rate and mean silent pause duration, but not mean syllable duration). However, articulatory skills explained the variance better in these measures (i.e. silent pause rate and duration) in L2 fluency (19% and 27%) than they did in L1 fluency (10% and %7, respectively). The authors concluded that 'having good articulatory skills may lead to more fluent speech in L1 and L2 speech, (at least) with respect to pausing, but not slower or faster articulation rate in semi-spontaneous speech'

(p. 14). They attributed the individual differences in speed fluency (i.e. articulation rate) to (lack of) L2 proficiency rather than variability in articulatory skills.

Huensch and Tracy-Ventura (2016)

In another recent study, Huensch and Tracy-Ventura (2016) also examined the influence of L1 fluency behaviour, cross-linguistic similarities/differences and L2 proficiency on L2 fluency behaviour over time. Two groups of English L1 speakers were recruited for the study, i.e. 24 L2 Spanish learners and 25 L2 French learners, and the data were collected through oral narrative tasks which the participants performed twice in their L2s (i.e. before and after 5 months of studying abroad) and once in their L1 (i.e. when they returned home). The participants' proficiency was measured through EITs in Spanish and French, and the same seven fluency measures used in De Jong et al. (2015) were employed. In order to allow for cross-linguistic comparisons of fluency between native speakers across the three languages, another two groups of L1 speakers were recruited, i.e. Spanish and French groups of native speakers. Huensch and Tracy-Ventura found significant differences between native speakers of English, Spanish and French for four fluency measures. English L1 speakers had a longer mean syllable duration than the other groups, and French L1 speakers had longer mean syllable duration than Spanish L1 group. L1 speakers also differed in their filled pause frequency and repair fluency; Spanish group produced fewer filled pauses and repetitions than the other groups, while English group made the highest number of corrections. Like other researchers, the authors argued that cross-linguistic differences (e.g. syllable structures of the given languages) might be affecting the fluency characteristics of the speakers.

Also, significant correlations were reported between L1 and L2 measures, before and after residing abroad, for mean syllable duration and number of silent pauses for both L2 groups (i.e. Spanish and French) and at both times. Huensch and Tracy-Ventura also observed that "L1 fluency behaviour, cross-linguistic differences, proficiency contributed differently to explaining L2 fluency behaviour prior to and during immersion" (p. 2). Before participants resided abroad, their L1 predicted only two L2 measures, i.e. mean syllable duration and number of silent pauses, and their proficiency scores predicted mean syllable duration and mean silent pause duration. However, after 5 months of residency, no influence of proficiency was observed while the contribution of L1 fluency in explaining L2 fluency behaviour was significant for all fluency measures. As for the influence of language group (i.e. L2 French and L2 Spanish), it predicted L2 fluency only for mean silent pause duration before residing abroad and the same fluency measures that were different between the groups (i.e. mean syllable duration, number of filled pauses, number of repetitions and number of corrections) after residing abroad. It was argued that the explanatory power of cross-linguistic differences and LP for L2 fluency changes over time, leaving more room for the effects of other factors, such as L1 fluency characteristics, on L2 fluency.

Peltonen (2018)

A final study to report here is Peltonen (2018), who examined L1 and L2 fluency connections from a mixed methods perspective. Working with 42 Finnish learners of English, she employed two picture description tasks to elicit data and a relatively wide range of fluency measures (i.e. 13 measures in total). The learners belonged to two school levels: ninth grade (Group1, B1 level at CEFR) and upper secondary school levels (Group2, B2 level at CEFR). Correlational analysis between L1 and L2 fluency measures indicated positive relationships for the majority of measures, with the strongest correlation reported for mean length of end-clause silent pauses (.68) and the weakest one reported for filled pause though still being moderate (.41). The majority of the measures also predicted L2 fluency at a significant level. The results overall suggested that L1 fluency might indeed be an important factor in explaining L2 fluency behaviour.

When she further compared the two groups to understand whether the links between L1 and L2 fluency could be affected by LP levels, she observed that Group2 was overall more fluent than Group1 in their L2 (i.e. Group2 having faster speech rate, longer mean length of run or shorter pauses), but the groups also differed in their L1 fluency. As the author acknowledges, however, proficiency groups in the study were not completely distinct, with large standard variations, and therefore the groups differences in terms of LP might have encouraged the large within-group variations in fluency performance.

2.4.1.4. Implications for Study 1 and Study 2

Having reviewed some of the previous cross-linguistic studies on L1 and L2 fluency behaviour (please see Table **2.1** for a summary of these), I will now discuss their shortcomings and what implications could be drawn from these for Study 1 and Study 2. Taken together, the results from this body of research suggest that one's fluency in L1 could be reflected in different ways in their L2 speech; e.g. through personal speaking styles or cross-linguistic effects. Cross-linguistic differences and/or L2 proficiency emerge as important factors potentially affecting L2 fluency behaviour. Regarding these, the findings from Derwing et al. (2009) and Huensch and Tracy-Ventura (2016) are particularly noteworthy. Significant correlations between L1 and L2 observed for both L1 groups in Derwing et al. (i.e. Mandarin and Slavic) were maintained only for Slavic group after residency abroad; this was argued to be due to 1) the effects of different amounts of proficiency gains in the two language groups, 2) cross-linguistic effects, i.e. closer relationship between Slavic and English than between Mandarin and English), or 3) the effects of both on L2 fluency. Likewise, in Huensch and Tracy-Ventura (2016), the predictive power of LP, TL group and L1 fluency over L2 fluency shifted over time; after 5 months of residency abroad, LP did not show any explanatory power over any of the L2 measures whereas L1 measures predicted all of the L2 measures and TL group predicted four measures (i.e. the same ones what were different across the languages in an earlier analysis). The findings from the two studies clearly suggest that several factors including LP levels and cross-linguistic effects might have mitigating effects on L1-L2 fluency associations.

The findings from De Jong et al. (2016)'s study are also of particular importance for the current studies in that they suggest that removing the effect of L1 fluency behaviour from L2 fluency might more accurately predict L2 proficiency. Indeed, the corrected L2 fluency measure of *mean syllable duration* better predicted L2 proficiency. This study also suggested that L1 fluency characteristics might be carried over to L2 speech through personal styles as did some other studies (e.g. De Jong & Mora, 2019; Huensch & Tracy-Ventura, 2016; Towell et al., 1996), and using such corrected measures might be more useful for L2 research and for assessment purposes.

Although the findings from the studies reviewed here make valuable contributions to our current understanding whether fluency is trait-like or state-like characteristic (Derwing et. al., 2009), and/or what fluency features are language-specific or universal (Riazantseva, 2001), several shortcomings and/or limitations can be identified in them. To start with, across these studies, there is no systematic approach in fluency measurement, which is a persisting issue in fluency studies mentioned earlier. Some studies focused on the investigation of pausing phenomena only (e.g. Riazantseva, 2001) or pausing and speed (e.g. De Jong & Mora, 2019; Derwing et al., 2009), while others employed a selection of measures to represent each fluency aspect (e.g. De Jong et al., 2015; Huensch & Tracy-Ventura et al., 2016; Peltonen, 2018). Equally important, there are variations in the operationalization of these aspects; while some investigated pauses of 100ms and above (e.g. Riazantseva, 2001), for others 250ms was taken as the starting point for pauses (e.g. De Jong et al., 2015; Huensch & Tracy-Ventura, 2015; Peltonen, 2018).

Furthermore, cross-linguistic studies examining L1 and L2 fluency patterns mostly worked with L2 learners at one proficiency level, e.g. intermediate to advanced group in De Jong et al. (2015), upper-intermediate group in De Jong and Mora (2019) or advanced group in Towell et al. (1996) or two levels (e.g. intermediate and advanced groups in Riazantseva (2001) or Peltonen (2018). Equally important, the assessment of proficiency in these studies are often unsystematic with some using a vocabulary test (e.g. De Jong et al., 2015) others using no test (e.g. Peltonen, 2018). However, given the crucial role of LP in L2 fluency development, knowing how learners behave across different LP levels, including low levels, is even more necessary. An examination of LP levels from a broader perspective will surely also contribute towards a better understanding of LP role in the relationship between L1 and L2 fluency. In fact, the finding in De Jong and Mora (2019) that L2 articulation rate, though correlated significantly with L1 fluency, could not be explained by variations in articulatory skills but L2 proficiency, is a further indication of the potential impact LP might have on the strength of the relationship between L1 and L2 fluency. To what extent LP affects this relationship is yet to be discovered.

Perhaps a more important issue in cross-linguistic studies is a methodological one. Majority of this body of research has not measured LP systematically; in these studies, LP assessment is usually based on one single test only partially relevant to the speaking skill; e.g. vocabulary test (De Jong & Mora, 2019; De Jong et al., 2015, Peltonen, 2018), an oral proficiency paper and pencil test or a grammar test (Riazantseva, 2001) or an EIT test (Huensch & Tracy-Ventura, 2016). These studies all seem limited in their assessment of proficiency as they cannot provide a complete and reliable measure of the participants' proficiency. While a vocabulary test or a grammar test provides an interesting insight into the participants' LP, it does not show much about their speaking skill. Similarly, using an EIT test only, would provide researchers with a narrow picture of the participants' speaking without any reference to their overall linguistic knowledge. The lack of a systematic approach in cross-linguistic fluency studies makes it even harder to compare results reliably or to make strong claims about the role of LP level in L1 and L2 fluency connections. It is well known that assessment of LP accurately and reliably is a crucial factor for the success of L2 research and/or teaching (whether for formative assessment, summative assessment or language certification) (Hulstijn, 2015; Leclercq & Edmonds, 2014; Tremblay, 2011), and a systematic LP assessment in L2 research not only contributes to capturing individual variations better amongst learners/participants, but it also gives insights into the linguistic development of learners (Tremblay, 2011). As such, if we are to properly understand the role of LP in the relationship between L1 and L2 fluency behaviours, a more systematic and comprehensive proficiency assessment seems crucial.

A final point regarding research on L1 and L2 fluency centres on the fact that the majority of them were performed with learners in a Study-Abroad context or participants who were immersed in TL community (Leonard & Shea, 2017). Given the research evidence that learners' fluency development benefits typically more from living in TL community (Freed et al., 2004; Huensch & Tracy-Ventura, 2017; Mora & Valls-Ferrer, 2012; Wright, 2013, 2018), it seems necessary to conduct studies in EFL contexts, where learners have not had extensive exposure to the TL or have not used the language for authentic communication purposes.

About the way L1 is reflected in L2 fluency, Derwing (2017) also notes that 'L1 may also serve as a proxy for other factors' (p. 251). She argues that socio-affective factors (e.g. time spent interacting in the TL, or certain learner traits or profiles) or motivational factors also provide explanations, to varying extents, for variability in fluency gains. Indeed, in Derwing et al (2008), the same group of learners in Derwing et al. (2009) were interviewed about the time they spent interacting in the TL (e.g. how often they had conversations in L2 or spent time listening to radio in English), and the findings suggested that Slavic group, who made more gains in L2 fluency development after 5 months of residency abroad in Derwing et al. (2009), deliberately exposed themselves more to the L2 whereas Mandarin group spent significantly less time doing so.

Yet, another factor affecting L2 fluency could be the type (i.e. picture-narrative or conversation) or nature (e.g. task design features) of the linguistic task used in different studies (Derwing, 2017, p. 252). The sources of the differences in such tasks results from how cognitively demanding they are (Foster & Tavakoli, 2009; Tavakoli & Foster, 2011). In addition to these, individual cognitive factors such as WMC (Skehan, 2015; Wen, Mota, & McNeill, 2015) or executive control skills involved (Shao, Roelofs, & Meyer, 2012) are amongst the contributing factors to differences in fluency performance.

To sum up, the findings from the existing research suggest that L1 and L2 fluency behaviour are associated with each other, and L2 fluency could be, at least to some extent, be predicted from L1 fluency behaviour or personal speaking styles[1]. In addition, a variety of factors, including typological features of L1 language, LP, amount of exposure to the TL (i.e. LoR), task design features, or cognitive factors such as WMC all contribute to

[1] In this research, the term 'personal speaking style' is used as a technical term to reflect L1 fluency behaviour. In line with the trait versus state issue (Derwing et al., 2009), the term 'personal speaking styles' is a relatively permanent characteristic which is specific to an individual. This includes speaking patterns observed in individuals' speech such as their tendency to speak slow or fast and use pauses, hesitations or repairs, and their tendency to use some lexical items more than others (e.g. low or high-frequency words).

development of fluency and therefore, might have a potential impact on the relationship between L1 and L2 fluency. However, the findings so far are neither conclusive, as they do not clearly show whether L1 and L2 fluency behaviours are related for typologically distant languages as well, nor generalizable to different research contexts such as EFL contexts. Equally important, we do not know to what extent LP or other similar factors might mediate the strength of this relationship. To this end, the findings need to be validated by more research. Working with L2 speakers belonging to more than one proficiency levels, more language pairs (i.e. L1-L2), including typologically distant ones, should be examined. LP should be assessed more systematically and comprehensively; i.e. by using more than one means rather than relying on one single tool and tapping into different linguistic knowledge types. Additionally, more studies are also needed to understand how such additional factors with a potential influence on L2 fluency mediate the strength of L1-L2 fluency relationships.

As such, Study 1 and Study 2 aim to address these shortcomings in the area in several ways. They aim to make a unique contribution to the filed by investigating the fluency links between two typologically distant languages, namely Turkish and English. They also seek to find out to what extent LP mediates this relationship, each by working with learners from *different proficiency levels* and in *different learning contexts* (i.e. ELF and study-abroad*)*. Both studies look at the speech performance from three proficiency levels, including low-levels as opposed to most research in the area which worked with participants belonging to only one or two proficiency levels (mostly higher levels). Additionally, LP is assessed more systematically, i.e. through the use of two proficiency tests of the OPT and EIT; each of these tests is assumed to measure different linguistic knowledge types; the former to measure declarative linguistic knowledge, and the latter to measure procedural linguistic knowledge. Both studies also aim to examine the potential impact of *the amount of exposure to the TL (through LoR), task design features (i.e. task structure)* and *individual variations in WMC* on the strength of these relationships. Although I acknowledge that other additional factors such as affective factors (e.g. willingness to interact in L2) or different task types (e.g. dialogue or conversation) also deserve an investigation, these are beyond the scope of the two studies reported in this thesis.

Table 2.1. A summary of the cross-linguistic studies examining L1 and L2 fluency with their methodologies used and the key findings

	Participants and LP levels/ groupings	Languages under investigation	Research context	Data elicitation method	Measures used	Key findings
Towell et al. 1996	12 learners of French at advanced level	L1 English L2 French	Study abroad (performance before and after 6 months of residency)	Film-retelling task (performed twice in L2, i.e. before and after residency abroad, and once in L1; i.e. a year later when they returned to home)	- speaking rate - phonation-time ratio - articulation rate - mean length of run - mean length of silent pauses	▪ After residing abroad, gains in L2 speed fluency- speech rate, articulation rate and mean length of run- were observed. ▪ Strong correlations between L1 and L2 fluency were reported for speech rate both before and after residing abroad.
Riazantseva (2001)	30 learners of English at intermediate advanced levels 20 native speakers of English (control group)	L1 Russian L1 English L2 English	Study abroad	A topic narrative and a cartoon description task, both: • consisting of different prompts • randomly assigned to participants to be performed in L1 and L2	- pause frequency - pause duration - pause location	▪ L1 groups differed significantly their pause locations; L1Russian speakers had longer pauses than L1 English speakers ▪ Advanced learners of English produced similar lengths of pauses as L1 English speakers ▪ Intermediate learners of English made longer pauses than L1 English speakers

						▪ Learners at both levels made more pauses in general and more within-constituent pauses than L1 English speakers.
Derwing et al. (2009)	16 native speakers of Mandarin 16 native speakers of Slavic languages All participants were at beginner levels of English at the start.	L1 Mandarin L1 Slavic languages (i.e. Russian and Ukrainian)	Immersion in Canada/ Study abroad (performance at three different times over a year: at 2 months, 10 months and 1 year of residency abroad)	Oral narrative tasks (performed once in L1 and three times in L2s)	Fluency ratings of trained judges on: - number of pauses - self-repetitions - false starts Other temporal measures in the performances: - number of pauses per second - speech rate - pruned speech syllables	▪ Fluency ratings in L1 and L2 correlated significantly for both L1 groups at 2-month time (stronger for Slavic group) ▪ At initial stages of L2 exposure, significant correlations were found between L1 and L2 fluency for all measures ▪ At later stages, significant correlations ere oberserved only for Slavic group
De Jong et al. (2015)	27 L1 speakers of English 24 L1 speakers of Turkish	L1 English L1 Turkish L2 Dutch	Study abroad	Eight speaking tasks (performed in both L1 and L2)	Corrected (for L1 fluency) and uncorrected measures: - mean syllable duration (i.e. inverse articulation rate)	▪ All measures could be predicted on the basis of L1 fluency measures (the amount of variance ranging from 21% to 57%) ▪ All measures predicted L2 proficiency significantly;

	All participants were learners of Dutch at intermediate to advanced levels.				- mean silent pause duration within ASU (ms) - mean silent duration between ASU (ms) - number of silent pauses /second - number of filled pauses /second - number of repetitions /second - number of corrections/second	corrected measure of mean syllable duration predicted LP more strongly than its uncorrected equivalence. ▪ Significant differences were observed for mean syllable duration and number of silent pauses between L1 two groups, i.e. Turkish and English.
De Jong and Mora (2019)	51 L1 speakers of Spanish L2 learners of English at upper-intermediate level	L1 Spanish L2 English	Foreign language school environment (EFL context)	To measure fluency, three speaking tasks: • formal descriptive task • a formal persuasive task • an informal persuasive task To measure articulatory skills, two other tasks • delayed picture-naming task	- mean syllable duration (i.e. inverse articulation rate) - number of silent pauses per min (speaking time) - mean duration of silent pauses	▪ Strong significant correlations were found between L1 and L2 fluency for all measures. ▪ Articulatory skills could explain only a small portion of the variability in L1 and L2 fluency (only for number of silent pauses and mean silent pause duration); larger variance could be explained by articulatory skills in L2 fluency than in L1 fluency.

				• speeded syllable production task		
Huensch and Tracy-Ventura (2016)	Two groups L1 English speakers: - 24 learners of Spanish - 25 learners of French All learners were at beginner levels at the start. Two control groups: - 18 L1 Spanish speakers - 10 L1 French speakers	L2 Spanish L2 French L1 English L1 Spanish L1 French	Study abroad (performance before and after 5 months of residency abroad)	Oral narrative tasks (performed once in L1 and twice in L2s)	- mean syllable duration (i.e. inverse articulation rate) - mean silent pause duration within ASU (ms) - mean silent duration between ASU (ms) - number of silent pauses /second - number of filled pauses /second - number of repetitions /second - number of corrections/second	▪ Significant differences were found between L1 groups (English, Spanish and French) for four fluency measures: mean syllable duration, number of filled pauses, number of repetitions and number of corrections. ▪ Significant correlations were found between L1 and L2 fluency for mean syllable duration and number of silent pauses for both L2 groups and at both times. ▪ LP predicted mean syllable duration and number of silent pauses only before residency abroad. ▪ L1 language predicted only mean syllable duration before residency abroad and four fluency measures (mean syllable duration, number of filled pauses, number of repetitions and number of

						corrections) after residency abroad.
Peltonen (2018)	42 L1 Finnish speakers All participants were learners of English at two school levels: ninth grade (B1) and upper school (B2) levels	L1 Finnish L2 English	Foreign language school environment (EFL context)	Two picture description tasks (performed in L1 and L2	Temporal measures: - speech rate - articulation rate - mean length of run - number of silent pauses - number of mid-clause silent pauses - number of clause-boundary silent pauses - mean length of silent pauses - mean length of mid-clause silent pauses - mean length of end-clause silent pauses Stalling mechanisms: - number of filled pauses - number of drawls - number of fillers - number of repetitions -	▪ Significant correlations between L1 and L2 for all measures, except three of them: mean length of mid-clause silent pauses, number of repetitions and number of fillers.

2.4.2. Lexical complexity

Lexical complexity is another important aspect of L2 oral performance. The vital role of lexis in L2 speech has been recognized in the literature in the sense that one's *lexical knowledge* tends to increase along with the LP level (De Clercq, 2015; Milton, 2013; Yu, 2010). Indeed, lexical complexity performance has been reported to be closely related with L2 proficiency in several studies (Iwashita et al., 2008; Révész et al., 2016; Yu, 2010; Zareva et al., 2005), and as such L2 researchers have widely employed vocabulary knowledge tests to measure the variability in participants' LP in their studies (e.g. De Jong et al., 2015). However, terms such as l*exical/word/vocabulary knowledge*, *lexical/word competence/skill*, *vocabulary/lexis use or* even *lexical complexity* itself seem to have been used in a wide range of ways; i.e. sometimes interchangeably and at some other times to mean different things (please refer to Henrik, 2013 for a detailed discussion of different terminology used). Below, I aim to clarify what lexical complexity means in the context of Study 2, and to that end, first I provide a brief discussion of what 'lexical knowledge' actually entails and how it is defined.

2.4.2.1. Word or lexical knowledge

From an acquisitional perspective, lexical or word knowledge has usually been represented through distinct levels; *breadth* or *size* of one's vocabulary, i.e. 'the quantity of words a person has some knowledge of' (Gonzalez-Fernandez & Schmitt, 2017) and *depth*, i.e. how well one knows about those words (Bulte, Housen, Pierrard, & Van Daele, 2008; Schmitt, 2010, 2014). Gonzalez-Fernandez and Schmitt (2017) explain that 'breadth has generally been conceptualized as knowledge of the form-meaning link of words. Depth, however, includes learning aspects such as the word class, collocations and grammatical functions, polysemous meanings, associations and constraints on use' (p. 284). These levels are not unrelated to each other nor do they develop in a parallel way; however, they contribute to one another (Li & Kirby, 2015; Schmitt, 2014). Gonzalez-Fernandez and Schmitt (2017) explain this with an example; as more words are learnt, more information (e.g. suffixes or prefixes) about the parts of those words will be placed in the lexicon and these will lead to an easier acquisition of the morphological aspects of vocabulary.

In L2 research, another useful distinction is made between *receptive* (or *passive*) and *producti*ve (or *active*) knowledge (Milton, 2009; Nation, 1990; Ongun, 2018). A learner's receptive knowledge involves the recognition of words when heard or read, and it is generally agreed that this knowledge of learners is greater than their productive knowledge. Productive knowledge, on the other hand, means these words can actually be accessed and used in speech or writing. Nation's (2001) framework of 'what is involved in knowing a word' (p. 27) also maintains this distinction between word knowledge types. Actually proposing to demonstrate all components of word knowledge in a more systematic and detailed way, the model classifies word knowledge into three areas; namely, form, meaning and use, and each of these are further divided into several sub-areas (see Nation, 2001, p. 27 for a detailed discussion of these). Table 2.2 below presents his framework; receptive versus productive distinction for these areas is mapped in the model with each of their sub-sections shown with R (receptive) and P (productive).

Table 2.2. Nation's framework of the dimensions involved knowing a word (2001, p. 27).

Form	Spoken	R	What does the word sound like?
		P	How is it pronounced?
	Written	R	What does the word look like?
		P	How is the word written and spelled?
	Word Parts	R	What parts are recognizable in this word?
		P	What word parts are needed to express meaning?
Meaning	Form and meaning	R	What meaning does this word form signal?
		P	What word form can be used to express this meaning?
	Concepts and referents	R	What is included in the concept?
		P	What items can the concept refer to?
	Associations	R	What other words does this word make us think of?
		P	What other words could we use instead of this one?
Use	Grammatical functions	R	In what patterns does the word occur?
		P	In what patterns must we use this word?
	Collocations	R	What words or types of word occur with this one?

		P	What words or types of word must we use with this one?
	Constraints on use	R	Where, when and how often can we use this word?
		P	Where, when and how often can we use this word?

Yet, another distinction is made between lexical knowledge and lexical competence (Laufer, 2003). *Lexical knowledge* is the information about the words stored in the mental lexicon, which could include grammatical properties of the words, their meanings and relations with other words (i.e. paradigmatic and syntagmatic relations) (ibid). *Lexical competence*, on the other hand, is not limited to the knowledge, whilst including it; competence also involves a person's willingness to use the knowledge, ability to access to it (i.e. speed and ease of the retrieval of the words from the lexicon), and strategic competence (i.e. strategies used by the speaker such as inferring meaning from context or negotiating of meaning). Laufer (2003) goes on to explain that if a person uses the word 'ask' instead of 'inquire' in their writing, this does not necessarily mean that the latter (i.e. inquire) is non-existent in their mental lexicon, but rather might indicate that the person has not put that knowledge into use or simply has not preferred to do so.

Regarding the distinctions between word types, Milton (2009) cautions that while such 'binary divisions' (e.g. breadth and depth, receptive and productive) might be useful in characterising differences in learner qualities, 'they do not really do justice to the complexity of word knowledge' (p. 13). He suggests that such terms might turn out to be quite ambiguous and cause confusion. While a vocabulary size test (i.e. breadth) might require a learner to provide translations of the words or some explanation of them, forming links with the meaning, a passive word recognition test might only necessitate the learner to recognize the words, without actually knowing or guessing what the words could mean. In such cases, though both are of the same test type (i.e. breadth), the former test is likely to result in an estimate of a smaller vocabulary size than the latter kind of test would. While a prolonged discussion of such different categorizations of word or lexical knowledge is not intended here, one could argue, from a processing perspective (i.e. the cognitive processes involved in speech), that the speed and ease of the lexical retrieval of items from the mental lexicon are closely related with fluency (Daller, Milton, & Treffers-Daller, 2007; Daller & Xue, 2007). Daller et al. (2007) note that fluency distinguishes what can be accessed from the lexicon and used from what is simply recognised and/or known; in this sense, fluency refers to the productive knowledge a speaker has.

2.4.2.2. Lexical complexity within CALF

If we return to the psycholinguistic views of speech production, discussed earlier in this chapter, we should recall that lexis constitutes a major construct in Levelt's model. In the CALF framework, lexical complexity or lexis is viewed more holistically, as one of the two parts of the construct of complexity, but at the same time, its interrelatedness to other aspects of the framework (e.g. fluency) and the importance of contextual factors (e.g. modality, register) are also stressed (Lahmann, Steinkrauss, & Schmid, 2019; Ortega, 2012; Skehan, 2009). In the CALF field, lexical complexity has generally been related with the characteristics of *vocabulary in use* which could be called lexical competence in Laufer (2003)'s terms; namely, *lexical diversity, lexical sophistication* and *lexical density* of the language produced in a given time (Read, 2000). Though speech characteristics could by no means be limited to these three dimensions - for example word relations or collocations also form a key part of it (Durrant & Schmitt, 2009) - the development of lexicon has generally been measured through some or all of these. The term, *lexical richness* (Read, 2000) or *lexical complexity* as will be referred to in Study 2 (De Clercq, 2015), is commonly used as a cover term for these sub-dimensions (i.e. lexical diversity, lexical sophistication and lexical density); however, as was the case with the fluency construct, definitions or operationalizations of these sub-dimensions also suffer from multiplicity. For example, what was called lexical diversity in Yu (2010) was referred to as lexical variety in Jarvis (2013). I will now discuss the definitions and the measurements of these sub-dimensions.

2.4.2.2.1. Lexical diversity

Lexical diversity (LD) is taken to be 'the variety of active vocabulary deployed by a speaker or writer' (Malvern & Richards, 2002). It is the ratio of unique words in a text or the speaker's ability to use or introduce new words (Skehan, 2009). LD is basically measured by calculating the percentage of different words in a text (i.e. *types*) to the total number of words used (i.e. *tokens*). The implication lying behind the use of this method in L2 research examining oral performance is that a greater variety of words in speech will correlate with a higher linguistic proficiency. Indeed, studies reported correlations between LD scores and proficiency test scores, e.g. correlations between LD scores and German C-test scores in Daller et al. (2003) or French C-test scores in Treffers-Daller (2013) or overall proficiency scores in Crossley, Cobb and McNamara (2013). These correlations suggest that LD measures could provide an estimate of general language ability.

However, type/token ratio (TTR) has been shown to be sensitive to text length in that as text length increases, the chances of using the same words within the text increase (Malvern, Richards, Chipere, & Duran, 2004). Several corrected measures have been proposed to address this shortcoming. Well-known examples of these include Guiraud Index (1954), D measure (Malvern & Richard, 2002; Malvern et al., 2004) or more recently, HD-D (McCarthy & Jarvis, 2007) or MLTD (McCarthy & Jarvis, 2010). All of these measures are essentially based on the calculation of TTR values in a given text, however, they use a series of different computations of TTR values. D measure, for example, is obtained through calculating TTRs in samples of different text lengths (i.e. typically between 35-50 tokens) and computing a random sampling of TTR curve, or HD-D measure takes a random sample of 42 words in a given text and for each type within the text, it calculates the probability of having any of its tokens in the random sample (McCarthy & Jarvis, 2010).

There are a number of studies that demonstrate the use of such corrected measures (Daller & Xue, 2007; Tidball & Treffers-Daller, 2008; Treffers-Daller, 2013; Treffers-Daller, Parslow, & Williams, 2018). Yet, while using these LD measures in research, several issues should be kept in mind. First, none of them have yet been able to solve the issue of text length sensitivity; this suggests, LD scores obtained through corrected measures, e.g. D or HD-D, still increase with text length. Concerning this issue, Treffers-Daller et al. (2018) suggest that simple count of *types* in a text, which could be considered as the most traditional method, could prove more successful in detecting within or between group differences than the use of such more recent/corrected measures which are based on some kind of complex formulae. Indeed, working with 179 adult learners of English coming from a variety of L1 backgrounds and a range of proficiency levels (i.e. B1 to C2 at CEFR), Treffers-Daller et al. examined the learners' essays which they wrote for an academic English language test. They employed both traditional measures (e.g. simple counts of types) and the new measures (e.g. Guiraud, D or HD-D) to find out which ones could better explain the variance amongst the CEFR levels. The findings demonstrated that the basic measures (i.e. simple count of the lemmatized types) were more successful in explaining the variance across different LP levels.

Another issue with the use of LD measures relates to the operationalization of the basic unit of analysis; i.e., what is counted as a different word (type) and what is not. Again, approaches in this regard differ; some researchers (Jarvis, 2002) take different inflected forms of a word (e.g. walking, walks, walked) as tokens of the same type while some others count all inflected forms as different types (Yu, 2010). However, caution is needed here since different units of counting are highly likely to result in different LD scores (Treffers-Daller & Korybski, 2015; Treffers-Daller et al., 2018). Such different approaches in the operationalization of basic units of counting may lead L2 researchers (teachers as well) to overestimate or underestimate learners' scores. Furthermore, lemmatized and non-lemmatized versions of the data are also likely to yield significantly different results (Treffers-Daller, 2013). Therefore, lemmatization and operationalization of basic unit of counting (i.e. types) are crucial factors that should be considered in examining LD.

It is also worth mentioning that LD measures do not consider formulaic language (i.e. lexical units consisting of more than one word) in the analysis, but rather they are based on the analysis of single words (Nation, 2001). In psycholinguistic studies, such lexical units are assumed to be stored holistically in the memory from where they are retrieved as chunks (Wood, 2010; Wray, 2002, 2012). It is suggested that the use of lexical units leads to more advanced language (Myles, 2012) and facilitates speech fluency since they are typically spoken more fluently, with no pause or hesitation within the unit (Conklin & Schmitt, 2008; Tavakoli, 2011; Tavakoli & Uchihara, 2019; Wray, 2002, 2012). The fact that formulaic language use is not considered in the analysis is a further limitation of LD measures, which deserves consideration in the analysis of spoken language.

Finally, in a relatively recent study by Jarvis (2013), it was suggested that LD is a perceptual phenomenon; i.e. it is dependent on the perceptions of human raters. He suggests that diversity should be considered together with six other text characteristics; namely, “variability”, “volume”, “evenness”, “rarity”, “dispersion” and “disparity” (p. 22-25), and thus that LD is actually only one aspect of the construct (i.e. richness). This is clearly a matter of construct validity since it implies that LD is in fact a multi-faceted construct in itself; examination of LD may not mean examining the whole construct but rather only one aspect of it. Therefore, new and improved ways of measuring LD is necessary (ibid).

Taken together, all these issues suggest that L2 studies examining LD are likely to have limitations as the aforementioned methodological problems remain unsolved. There seems to be no standard unit of counting used in LD measures, nor is there one single measure that is adequate to represent the underlying construct truly and fully. However, although each method discussed so far has its own limitations with text length sensitivity being the most common problem, such problems do not invalidate the use of LD measures. There is considerable research evidence suggesting that they could be used as an overall measure of language development (e.g. Malvern et al., 2002) and/or to contribute to our understanding of oral performance in CALF studies (Bulte & Housen, 2012). Although not offering complete solutions to the existing problems, a few points could be considered when using LD measures. For example, the text length should be kept constant across different texts (McCarthy & Jarvis, 2007; Treffers-Daller et al. 2018), and clear explanations should be provided on what a type/different word is and how it is counted. Considering such points could increase the comparability of LD scores across studies. Finally, using both traditional measures (i.e. simple counts of the types) and more recent ones (e.g. Guiraud, D, HD-D or MLTD) rather than depending upon one single measure could provide a better measurement of LD.

2.4.2.2.2. Lexical sophistication

While LD could be related to the size of lexicon, *lexical sophistication* (LS) concerns to the depth of knowledge (Bulte & Housen, 2012), i.e. what proportion of the words used in a text are difficult or low-frequency when compared to a target corpus. Measures of LS calculate how many words (defined as low-frequency) are used within a given text based on frequency lists in that language. Similar to LD measures, the assumption underlying LS measures is that appearance of low-frequency words in a text or speech sample will relate to a higher proficiency level because low-frequency words are considered to be advanced or more difficult and therefore learnt at more advanced stages of acquisition (Laufer & Nation, 1999; Lindqvist, Bardel, & Gudmundson, 2011). However, LS measurement, too, is not unproblematic since existing word frequency lists could be quite context-dependent or be affected by such factors as genre or modality (Johansson, 2009; Lindqvist, 2010). Additionally, relying solely on word frequency-based measures in examining lexical complexity might bring forth other problems. For instance, language-specific factors (e.g. cognateness) could influence the scores in the sense that some cognates existing in a speaker's L1 could be more readily available to be used in their L2 (Bardel, Gudmundson, & Lindqvist, 2012); if such L1 cognates happen to be low-frequency words in the L2, this might lead to a conclusion that the speaker has a higher lexical sophistication.

2.4.2.2.3. Lexical density

The third and last aspect of lexical complexity is *lexical density,* which refers to the proportion of content words to function words used within a text (Bulte & Housen, 2012). This aspect pertains to information packaging of a text; in other words, a text containing a high proportion of content words (e.g. nouns, verbs) is considered to be denser in information than a text with a high proportion of function words such as prepositions, conjunctions, count nouns, pronouns, interjections (Johansson, 2009, p. 64). One important issue about lexical density, as noted by Johansson (2009), is that it is “highly language dependent” (p. 66); for example, a text produced in a given language would consist of a higher number of content words, if the language has more bound morphemes, than a text produced in another language, e.g. a typologically different one. If we take Turkish and English as examples, a word in Turkish (e.g. ‘evinde’) could equal to (in meaning) one word in English (e.g. ‘in his/her house’). This suggests the number of function words could be different in two texts of the same length produced in these two languages. For this reason, a comparison of typologically different languages (Turkish and English in this case) in terms of lexical density may be misleading.

Johansson (2009) also argues that ‘it is often necessary to re-use several function words in order to produce one (new) lexical word’ (p. 83), therefore, as a text becomes longer, more function words are likely to be used leading to a lower lexical density and TTR value as well. This issue actually relates to the text length sensitivity mentioned above. Related to these issues, Pallotti (2015) argues that we may have no objective grounds to locate lexical density into the wider context of lexical complexity, which clearly leaves the place of lexical density open to debate.

In the above discussion, I have so far attempted to explain the importance of lexical complexity aspect of oral performance in L2 research and how researchers situate it in the framework of CALF as one part of the complexity construct. I have mentioned the multi-faceted nature of lexical complexity, described each of its aspects, namely, LD, LS and lexical density, and discussed some of the existing measures for each aspect and their limitations. Given that Study 2 is interested in examining lexical complexity in L1 and L2, I now move on to a discussion of cross-linguistic studies in this regard that are most relevant to the current research.

2.4.2.3. Cross-linguistic research on lexical complexity

Given that lexical development is a key part of linguistic proficiency across different language domains such as reading, listening, writing or speaking, it is not surprising that lexical knowledge development across different languages has also been of great interest to L2 researchers. To this end, SLA research has pointed to cross-linguistically variable trajectories of lexical knowledge development; whether across different L1s, e.g. in Berman and Verhoeven (2002) where less LD increase was observed in Hebrew and French than other languages examined (Dutch, English, Spanish, Swedish and Icelandic), or across L2s, e.g. in Mardsen and David (2008) where higher LD values were found in Spanish than in French. Comparative research on L1 and L2 lexical complexity of the same learners is very rare, though with a few exceptions (e.g. Daller et al., 2003; De Clerq, 2015).

One example of such comparative studies is Daller et al. (2003), who examined lexical complexity in spontaneous speech of two groups of Turkish-German bilinguals, with one group consisting of Turkish-dominant bilinguals and the other group consisting of German-dominant bilinguals. The speech samples were elicited through a picture description task both in German and Turkish, and Guiraud index and advanced Guiraud Index were used to measure LD and LS, respectively. The analysis of the participants' language production in German for LS was based on a well-known frequency list in German (i.e. Oehler, 1983). However, since there was no such reliable list for the analysis of LS in Turkish, the authors opted to use an alternative way: they asked a group of experienced Turkish teachers to indicate which words in Turkish speech samples fell into basic and advanced categories. The results showed no significant differences in LD scores across languages, however, LS scores (i.e. advanced Guiraud Index scores) were higher in Turkish than in German for both groups, independently of their dominance in one language. The authors interpreted the findings with structural differences in German and Turkish. They argued that there are more high-frequency function words in German than in Turkish and this might have led to an overall lower LS score in German. However, one alternative explanation for this result might be explained in the light of different approaches adopted in the LS measurement in the two languages, i.e. using frequency lists in German and the teachers' judgements in Turkish might have confounded the results.

In another cross-linguistic study, Dewaele and Pavlenko (2003) explored whether culture and language affected LD and productivity in the speech of 258 participants, who were divided into five groups: two monolingual control groups (American and Russian) and three Russian-English bilinguals groups (speaking Russian as L1, speaking English as L2 (ESL) and speaking English as an EFL). They used Uber index, which is an arithmetical transformation of TTR, to measure LD in the participants' speech. The results demonstrated significant differences in LD scores across the control groups; the scores were significantly higher for Russian monolingual and bilingual (speaking Russian as L1) groups than American monolingual group. Interestingly, the Russian bilinguals speaking English as an ESL approximated their speech to the American monolingual group, and Russian bilinguals speaking English as an EFL produced lower LD scores than the American monolingual group. The authors provided several possible explanations for their results. They posited that differences in monolingual groups could be attributed to the typological differences across the languages (i.e. Russian and English), and/or to cultural requirements such as different ways of conceptualizations in the two languages and cultures. As for the finding that Russian users of English as an L2 approximated their performance to the American group, they interpreted it with Pavlenko's (1999) cultural competence; in the second language socialization process, this group may have conceptualized new linguistic concepts by transforming the existing conceptual representations in their mind, and as such, the influence of their L2 could have been reflected in their L1s. Finally, the authors also mentioned other possible factors, such as 'unfamiliarity with the cultural requirements' or low L2 proficiency level, which may have led to low LD scores in the speech of Russian users of EFL (p. 135).

I have so far argued that lexical or word knowledge has been characterised and categorised in different ways in L2 research. Although scope and space limitations have not allowed me to go into a prolonged discussion of such categorizations here, I have made it clear that the interest of Study 2 lies in lexical complexity as one form of linguistic complexity in the CALF framework. Adopting a psycholinguistic perspective, lexical complexity is taken in the study as one aspect of L2 oral performance, which has been closely related with other aspects (e.g. fluency) and L2 proficiency and is examined in terms of the speech characteristics such as LD, LS or lexical density. Later in Chapter 5, I will discuss existing issues in the measurement of this construct and what factors could influence the lexical performance in this sense.

2.5. Individual learner variables

As was discussed earlier, any factor that influences L2 fluency such as individual learner differences could potentially impact the relationship between L1 and L2 fluency behaviours. This research is interested in investigating the mediating roles of three individual learner variables; namely, L2 proficiency, WMC and LoR. Below, I explain the importance of each of these and discuss how they can be influential on L1-L2 fluency links.

2.5.1. L2 proficiency

Language proficiency (LP) is one of the fundamental variables that distinguish L2 speakers from one another in L2 research. Although a fundamental concept in many L2 studies, LP is often taken for granted in most studies failing to define and operationalize it systematically. The definition used in the current project draws on Hulstijn's (2015) work in which proficiency is conceptualised as "knowledge of language and the ability to access, retrieve and use that knowledge in listening, speaking, reading and writing" (p. 21). The definition brings our attention to two dimensions of LP; *the linguistic knowledge* across different domains and modalities and *the ability to use (including the access and retrieval)* that knowledge. Yet, in SLA research, the nature of this knowledge and its characteristics have been at the centre of debate. Regarding the types of linguistic knowledge, two important dichotomies bound: explicit vs implicit knowledge and declarative vs procedural knowledge.

To start with the first dichotomy, explicit knowledge refers to the knowledge which is obtained consciously and through analysis while implicit learning is intuitive, gained without consciousness or awareness (DeKeyser, 2003; R. Ellis, 2005; Hulstijn, 2005, 2015). The 'awareness' or 'attention' seem to be the key criterion in distinguishing the two types of knowledge (R. Ellis, 2005, p. 151-153); however, these types of knowledge are acquired or whether they are related to each other or are distinct constructs (i.e. the so-called interface issue, see N.C. Ellis, 2005 for a detailed discussion) have been the central topics in this field. Researchers have taken different theoretical positions to these ends; some argued that the acquisition and use of the two types of knowledge require different mechanisms, rejecting any interaction between the two (Hulstijn, 2002; Krashen, 1985) while some others, especially in the field of cognitive sciences, acknowledged the role of explicit knowledge in facilitating the acquisition of implicit knowledge through practice (DeKeyser, 1998, 2015) (though disagreements also exist on the nature of this practice).

When it comes to the second dichotomy, a distinction is made between declarative and procedural knowledge. The former is defined as the 'knowledge of facts (semantic memory) and events (episodic memory) usually consciously accessible and often verbalizable, but not necessarily' whereas the latter refers to the knowledge that can only be put into use (DeKeyser, 2017). DeKeyser (2017) argues, however, that the distinction between these two does not exactly match the other distinction between explicit and implicit knowledge. Explicit knowledge is usually considered to be of declarative nature (Paradis, 2009; Ullman, 2015); yet, declarative knowledge, while mostly explicit, could be implicit as well (DeKeyser, 2017, p. 16). In this sense, declarative knowledge mostly consists of explicit knowledge but could not be limited to it.

DeKeyser (2017) further explains that proceduralization (i.e. developing the procedural knowledge) is achieved only when learners are able to put their knowledge into use by acting upon their declarative knowledge repeatedly through practice. As a result of this frequent practice, learners become better in this process, by 'using it (their knowledge) more correctly, more easily and more frequently, in a variety of contexts' (p. 16). He maintains that although some people call this process 'automatization', in effect, what happens is that learners develop procedural knowledge (ibid). Automatization is achieved only after the development of the procedural knowledge and takes a very long time for most learners. Therefore, what is suggested here is that explicit knowledge transfers into automatized explicit knowledge.

Related to this, Suzuki and DeKeyser (2019) further argue that automatized explicit knowledge, although accessed quickly (and easily), should still be distinguished from implicit knowledge, i.e. the knowledge characterized by a lack of awareness. They advocate that the two are distinct constructs based on the awareness factor, and as such they should be measured separately. While a detailed discussion of these types of linguistic knowledge is not intended here, I should note that in Study 1 and Study 2, LP is examined in line with the distinction of declarative versus procedural knowledge only and neither study aims to distinguish between implicit and automatized explicit knowledge. In Hulstjin (2015)'s definition cited above, LP is characterized by two aspects, i.e. *the linguistic knowledge* and *the ability (or skills) to use* it. DeKeyser (2017) posits that 'skill is a form of knowledge' and thus declarative and procedural knowledge are sometimes called 'knowledge that' and knowledge how' (p. 16). This suggests that the two knowledge types, i.e. declarative and procedural, could arguably be matched to the two LP aspects in the definition, i.e. the linguistic knowledge and the ability (or skills) to use it, at least for the purposes of Study 1 and Study 2. As such, although limiting LP to these two aspects could have its own limitations such as not obtaining a full picture of one's linguistic knowledge, distinguishing between declarative and procedural knowledge seems well justified for LP assessment in these studies.

L2 studies have commonly examined LP in terms of either of the two knowledge types, i.e. declarative or procedural, by employing a variety of tools and methods to characterize proficiency levels of their participants or to describe the variations in participants' language development. This bring us to the next point of discussion, i.e. LP measurement in L2 studies.

2.5.1.1. Measuring L2 proficiency

Proficiency assessment has been widely used in L2 research to describe differences in dependent variables under examination or to group learners into different levels based on their overall L2 proficiency or on a specific skill (e.g. vocabulary size or fluency in a task). To achieve these ends, a variety of methods and tools have been employed across studies with some using scores or certificates obtained from objective assessments (e.g. IELTS, TOEFL, Cambridge Certificate of Proficiency), shorter and/or simplified versions of standardized tests (e.g. Oxford Placement Test, Cambridge University Test) or tests assumed to measure global proficiency such as vocabulary tests (e.g. cloze-tests or vocabulary size tests) or elicited imitation tasks (see Hulstjin, 2015 and Tremblay, 2011 for a review of different tests used across studies). This wide range of different methods utilised as a reference to LP levels of participants or to explain variations in the dependent variables points to a clear lack of uniformity and systematic approach in LP assessment.

Furthermore, regarding the way LP levels are defined and used across studies, Hulstjin (2012) argues that there is currently 'no psycholinguistic or sociolinguistic theory' on which LP levels could be based in studies (p. 429). Therefore, test performance is commonly expressed as scores or interpreted using scales such as CEFR. He maintains that level labels (e.g. beginner, intermediate, etc) or labels used by companies (e.g. A1, A2, B1 in CEFR) are commonly adopted in L2 research to describe individuals' LP. In fact, the CEFR levels have perhaps been the most influential references for LP levels in L2 practices, yet, as he argues elsewhere, CEFR levels, 'fail to consistently distinguish between L2 development and L2 proficiency' at higher levels (i.e. B2, C1 and C2) (Hulstijn, 2011). Further, in his study, a close inspection of the scales for higher LP levels at CEFR, revealed that performance at these levels required high language cognition skills, i.e. those that distinguish native speakers and involves the production of more complex, and often longer, language in terms of the lexis and grammar). This suggests that higher levels may never be attainable for most native speakers, nor for L2 learners unless they are functioning in higher professions or have high intellectual capacities. As also acknowledged in these studies (i.e. Hulstjin, 2011, 2012), however, LP levels are commonly (and maybe inevitably) used in L2 practices (e.g. research, teaching or assessment) for practical reasons, despite the shortcomings of scales such as CEFR.

This brief discussion points to two important conclusions about the way LP is assessed in L2 research. First, unfortunately, there has been neither a systematic approach nor uniformity, in LP assessment across studies. Given that LP has a direct influence on L2 learners' performance (e.g. Tremblay, 2011), this individual variable should be described well and be measured as accurately and systematically as possible. Second, describing LP levels, despite a lack of theory, seems necessary for practical purposes; e.g. to assign selected participants into different LP groups with the aim of describing the nature of the sample or explaining individual differences in language development, as well as for comparability of results across studies.

Following from these, I should note that LP assessment, both in Study 1 and Study 2, is based on the measurement of the two underlying constructs of declarative and procedural knowledge, as opposed to most research examining either of the two knowledge types. As such, the assessment tools are selected to represent each of them (mostly though not exclusively); namely, standardized tests of OPT to measure declarative knowledge and EIT to measure procedural knowledge. It is assumed that using two tests to measure the two kinds of knowledge will not only provide us with a broader perspective of LP, but it will allow for a more valid interpretation of the LP effect as a mediator on the strength of L1 and L2 fluency relationship. The CEFR scales are also employed as a reference to classify the participants into different LP levels. Further discussion of these will be provided in the Methodology Chapter (Section 3.5.2).

2.5.1.2. Language proficiency and L2 fluency

As discussed earlier, LP has been directly related with L2 development and L2 acquisition in general, and with the fluency aspect of oral performance in particular. Since Study 1 and Study 2 are interested in examining the potential impact of LP as an individual factor on the strength of the relationship between L1 and L2 fluency behaviours, a few words must be said about its role in understanding this relationship. From a psycholinguistic perspective, what distinguishes L1 and L2 speech production is the nature of processing mechanisms involved in each. L1 speech production process largely relies on incremental, parallel and automatic processing (Levelt, 1989; 1999), which ultimately helps make speech more fluent with little undue hesitation or interruption. On the other hand, in L2 speech production, parallel processing is disrupted by challenges faced by the speaker (e.g. lexical access and retrieval difficulties or monitoring process during and/or after the production), mainly due to L2 speakers' incomplete linguistic knowledge (Kormos, 2006, 2011; Skehan, 2015). Therefore, L2 speech is assumed to be serial, mostly operating on conscious and controlled mechanisms, especially at lower levels of proficiency. Such difficulties encountered in the L2 case are often reflected in the overt speech with signs of disfluency such as frequent pauses, hesitations, self-corrections or slower speech. However, as the LP increases when learners expand their linguistic knowledge, they increase their automaticity and parallel processing begins to surface, making their L2 speech like L1 in that it becomes more automatic and more fluent. From this viewpoint, one can hypothesize that at higher LP levels, one might expect a stronger relationship between L1 and L2 fluency behaviour largely because at higher levels L2 speech is more likely to resemble L1 speech. In contrast, at lower LP levels, since the challenges in the production process persist, though to varying degrees, the relationship between the two might be blurred (i.e. becoming less related with each other).

Whether LP impacts the strength of the relationship between L1 and L2 fluency behaviours is what both Study 1 and Study 2 aim to explore. While Study 1 is interested in the LP effects for learners at *lower-levels* (A2, B1 and B2 at CEFR) in an EFL context, Study 2 explores this for learners at *higher-levels* (B1, B2 and C1at CEFR) in a study-abroad context. It is important to note here that neither of the studies aim to compare LP in L1and L2 nor measure L1 proficiency. At this point, I find it necessary to say a few words about one's proficiency in their first language. Hulstjin (2015) notes that L1 speakers differ to a large extent in their language proficiency 'as a function of their individual attributes such as age, intelligence, knowledge of the world, level of education, professional career and leisure-time activities' (p. 28). However, he also adds that although native speakers differ mostly in their higher language cognition abilities (e.g. producing more complex and often longer, language in terms of the lexis and grammar), all speakers of L1 share basic language cognition abilities (i.e. implicit linguistic knowledge in different domains such as phonology, phonetics or morphology, explicit linguistic knowledge in the lexical domain and the automaticity in the use of these knowledge types). Neither study presented here aim to distinguish between higher and basic language cognition abilities of L1 speakers, but rather both studies seek to understand whether L1 fluency behaviour is mirrored in L2 fluency at an individual level regardless of the participants' level of L1 proficiency (i.e. whether each individual's L1 speaking styles are carried over to their L2 speech). Therefore, it was not deemed necessary to measure L1 proficiency in either study. Still, every effort was made to control for participants' background in terms of their education levels, professions and ages. The participants in both studies belong to similar educational backgrounds with all of them having completed at least a bachelor degree in Turkey, similar professions with all of them studying at postgraduate degrees in the UK and a similar age group (22-35).

2.5.2. Working memory capacity

It is widely accepted that learners' cognitive abilities play a large role in the acquisition of both L1 and L2. Among these, WMC has been one of the most important and widely examined components of human cognition (Baddeley, 2017; Gathercole, 2007; Wen et al., 2015). Working memory (WM) has been investigated in L2 research for its direct and indirect connections with language development, language processing and production (Wen & Li, 2019). WM is defined as a limited capacity system, responsible for holding the information temporarily and processing it when speakers are engaged in performing a complex task, such as speaking or problem-solving (Baddeley, 2003, 2015, 2017; Baddeley et al., 2009). It is suggested that performance in complex tasks requires a person to simultaneously maintain and manipulate the information (by controlling their attentional resources), and in fact, such tasks could range from anything from activities that we are engaged in our daily lives such as formulating a plan or following directions to language-related activities such as vocabulary learning, reading, or speech planning. Therefore, unsurprisingly WM has been brought to the attention of L2 researchers over the last few decades.

In L2 research, WM has been examined as an individual cognitive factor (i.e. as psycholinguistic constraint), and its effects have been implicated in a variety of L2 domains; e.g. vocabulary learning (Speciale, Ellis, & Bywater, 2004), language proficiency (Hummel, 2009; Kormos & Safar, 2008), grammatical development (French & O'Brien, 2008) fluency development (O'Brien et al., 2007) or lexical complexity (Gilabert & Munoz, 2010). The rationale for incorporating WM in this line of research came from the central assumption that if WM is a learner characteristic, individual differences in this limited capacity could explain some of the variability among learners' performances in complex tasks. In this sense, WM is sometimes seen as one aspect of language aptitude, i.e. a stable learner characteristic comprising several other cognitive abilities (e.g. phonetic coding ability or grammatical sensitivity) (Carroll, 1981; Skehan, 1998). Although views on whether WM is to be considered as one potential component of language aptitude or a separate cognitive learner factor (Dörnyei, 2014; Hummel, 2009) are not conclusive, there have been various attempts at defining the structure and functions of the construct in the literature. A full review of these is not the objective here (Miyake & Shah, 1999); however, it would be necessary to bring forward at least one to contribute to a better understanding of WM effects in language learning contexts in general and L2 oral performance in particular. I will now review the most influential WM model by Baddeley and Hitch (1974), which was later revised in Baddeley (2000).

2.5.2.1. The structure and functions of working memory

A proper understanding of how WM is influential on language development and performance requires one to be clear about how human memory functions during cognitive activities, including but not limited to language-related tasks. To that end, Baddeley and Hitch (1974) proposed a WM model for researchers. The model originally consisted of two short-term stores (i.e. slave systems) - *phonological loop* and *visuo-spatial sketchpad* - and a *central executive system*. The two stores are devoted to storing domain-specific information; the phonological loop stores phonological and verbal information and the visuo-spatial sketchpad stores and manipulates visual and spatial information. The former is postulated to further involve a *phonological store*, which maintains the information for very brief moments (a few seconds), and *an articulatory rehearsal component*, which keeps the information active through sub-vocal speech.

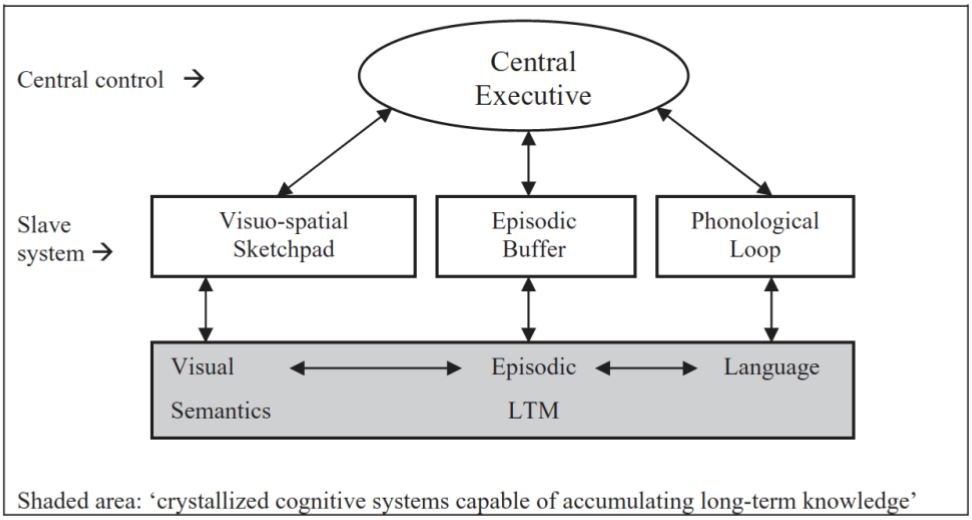


Figure 2.3. The revised multicomponent model of WM (Baddeley, 2000)

In a later version of the model (Baddeley, 2000; 2003), a fourth component was added to the model as another storage domain: *episodic buffer*, which serves as the gateway to long-term memory (LTM) (Figure 2.3). Baddeley (2015) defines episodic buffer as 'a multidimensional storage system, capable of combining information from the visuospatial and verbal subsystems and linking it with further information from the perception and long-term memory' (p.20). He suggests that it integrates different types of knowledge from the two stores with LTM and therefore is supposed to be responsible for the storage role. Finally, the central executive system coordinates the flow of information to the slave systems and to the LTM. It is also in charge of controlling and allocating attentional resources available during performance in tasks, e.g. by maintaining the focus, diverting attention or blocking the irrelevant information (Baddeley et al., 2009; Gathercole, 2007). In these aspects, it could be suggested that the central executive is at the heart of the WM system and has a key role there (Baddeley, 2003, 2015).

The above description brings our attention to two key functions of this hypothetical cognitive system: 1) holding the information in the memory and 2) manipulating that information while regulating the attentional resources during a complex cognitive activity. These two functions pertain to *storage* and *processing* elements of the system, respectively. It is assumed that these elements operate on a trade-off; when more information is stored, less capacity is left for processing. A third one could be that WM also functions as a gateway to the LTM in that it works in coordination with LTM with two-way flow of information between the LTM and the other components of the system to support language and cognition (Baddeley, 2015) Also, whether the storage and processing elements of the system are independent or interdependent to each other remains an issue yet to be established with certainty (Fiedman & Miyake, 2004; Gilabert & Munoz, 2010).

What is well-established, however, is that WM is limited in capacity and attentional sources. This means that our memories have constraints in how much information it can effectively maintain, process and integrate all at the same time. Gathercole (2007) illustrates this with an example: the task of calculating two two-digit numbers requires one to keep the numbers in the memory, but at the same time to execute the rules for a solution, which need to be retrieved from the LTM. In fact, several calculations are carried out in the process before a final solution is reached successfully. On the other hand, we are less likely to calculate two three-digit numbers due to the limits in the storage and attentional sources needed for task completion. Since the information is maintained only temporarily in the memory, when processing demands increase in a given task, attention is diverted from the maintenance activity and the information is lost (unless the person uses some kind of strategy to keep it, e.g. repetition) (ibid). Therefore, in such cognitively demanding tasks or complex tasks, a selective attention is needed to maintain the information and the focus of the task and to block any interference with task completion (Baddeley et al., 2009). It is important to remember that it is *the central executive system* which supervises and coordinates these activities (e.g. by switching the focus, dividing and directing the attention, linking the information with the slave systems and the LTM, suppressing the interference or irrelevant information) and the limited capacity (e.g. by sharing its limited capacity with the slave systems or allocating more capacity to them). In fact, regarding the ability of the central executive in blocking interference or irrelevant input could be assumed to be particularly important in speech production in the sense that it holds off the interference from L1 lexicon or previously learnt L2 words which sound similar) (N. C. Ellis, 2006). This suggests that more attentional sources are needed for L2 learners at lower-level learners to block these associations, and that those who have more attentional resources or can use their sources better are likely to produce more fluent language.

Researchers in L1 and L2 studies have increasingly recognized WM roles in L1 and L2 acquisition equally. In fact, WM might have even a greater role in L2 learning and performance; since L2 learners have to be content with a much smaller lexicon and a restricted L2 grammar system, there might be much greater demands for WM resources when language comprehension or production is involved (Skehan, 2015). As such, L2 research on WM has largely centred on the two WM components; phonological store, for its responsibility in storing verbal material and central executive, for its responsibility in regulating the storage and attentional resources.

In language learning and use, the manipulation of the verbal information (i.e. information stored in the phonological store) is of central importance (Baddeley et al., 2009). For this reason, one line of research has set out to investigate the role of phonological WM (PWM). PWM is assumed to help learners acquire chunks or word sequences and ultimately transfer them into the LTM (N. C. Ellis, 2012) and therefore, limits in PWM could be correlated with L2 development. Indeed, research has demonstrated close links between PWM and L2 specific domains of vocabulary learning (N. C. Ellis, 1996, 2012), formulaic language (Foster, 2013; Foster, Bolibaugh, & Kotula, 2014) and grammatical development (O'Brien, Segalowitz, Collentine, & Freed, 2006; O'Brien et al., 2007). This suggests that learners with larger WM capacity were found to be better across these domains. In another line of research, WM effects have been investigated in L2 comprehension, L2 interaction and L2 production across different modalities. The assumption is that for learners to comprehend a reading text, for example, they are required to hold several previous materials in memory and to simultaneously process the following sentences (Daneman & Carpenter, 1980). To this end, studies have reported positive correlations between WM and L2 domains: listening (Miki, 2012), reading comprehension (Alptekin & Erçetin, 2009, 2011), interaction (Révész, 2012) and performance (Ahmadian, 2012b, 2013; Guara-Tavares, 2008). Individual differences in WMC have therefore an explanatory power, at least to some extent, for variations in language development and performance across different L2 domains.

It is also worth mentioning that recently, Wen (2015) has put forth 'a more principled approach' to the investigation of WM in L2 contexts (p. 51), proposing a new model called Phonological/ Executive (P/E) Model. Wen (2015), among others e.g. Baddeley (2015; 2017) and R. Ellis (2005), argues that a model integrating WM in SLA studies is needed so that L2 research could focus on the key components (i.e. PWM and EWM) and their functions as these have been shown to be directly related with language learning and processing. He defines the P/E model, or the so-called 'integrated WM-SLA framework, as 'a limited capacity of multiple mechanisms and processes implicated in L2 domains and activities' (p. 52). His definition clearly echoes the unifying WM characteristics mentioned above, i.e. having a limited capacity and multiple components together with their devoted functions. Wen further postulates that these WM components are also interacting 'bidirectionally' with the LTM, i.e. the place where the representations of different knowledge types (e.g. L1 lexicon and grammar and L2 declarative and procedural knowledge) exist. Also, the components of the episodic buffer and visuo-spatial sketchpad are excluded from the model.

Finally, the P/E model attempts to offer a more systematic approach for the use of WM span tasks (i.e. simple span and complex tasks) in SLA research. In fact, it is postulated that WM could be best defined as what complex (i.e. involving both storage and processing elements) and simple span tasks (i.e involving only storage element) measure as separate measures are needed to assess the WM components. This brings forth the question of how WM or its components are measured in L2 research. I will discuss this next.

2.5.2.2. How to measure working memory capacity

In order to reflect the individual differences in the extent WMC is efficiently used by learners, two types of WMC tasks have been widely employed by researchers: 1) *simple span* tasks to measure individuals' ability to store and rehearse input (i.e. measuring the storage capacity only) and 2) *complex span* tasks to measure individuals' ability to store input while simultaneously engaging in additional processing (i.e. measuring both storage and processing capacities) (Juffs & Harrington, 2011; Linck, Osthus, Koeth, & Bunting, 2014). Simple span tasks includes forward digit/word span tasks or non-word span tasks, which involve immediate recall of a set of words, digits (numbers) or visuals after these are presented to individuals for brief amounts of time, usually in sets of increasing size and either in oral or written form (Baddeley et al., 2009). The most commonly used examples of simple span tasks include non-word repetition tests where the non-words (i.e. those that are meaningless and do not exist) conform to the phonetic rules of a given language (Mackey & Sachs, 2012). The assumption underlying the task is that previous knowledge of the language, which could influence task performance, could be avoided. A similar one is serial non-word recognition task (SNRT) (O'Brien et al., 2006) which is again comprised of non-words but unlike to common simple span tasks which involve some form of production, SNRT does require a test-taker only to recognize the order of the non-words.

Complex span tasks, on the other hand, require an individual to actively process the information while maintaining and recalling a set of letters, numbers, or visual objects. Some examples include reading and listening span tasks (Daneman & Carpenter, 1980; Van den Noort, Bosch, & Hugdahl, 2006) and Backward digit span tests (Awwad, 2017; Awwad, Tavakoli, & Wright, 2017). In a reading span task, for instance, a string of sentences are presented to individuals and they are asked to first determine if the sentence is acceptable and then to recall a target word for each of the sentences upon the completion of the task. Other complex span tasks include different versions of reading span task, e.g. listening span (Goo, 2010) and speaking span (Finardi & Weissheimer, 2009), and operation-span (O-span) task and its variants (Weitz, O'Shea, Zook, & Needham, 2011). In O-span tasks, individuals are typically given a series of mathematical problems to deal with, which are followed by an L2 word (e.g. high-frequency nouns), and they are then required to remember the words in order.

The measurement of WM also suffers from a lack of consensus among researchers, regarding the tools employed and different functions tested. Wen (2015) argues that confusion surrounds particularly the use of terms; WM measurement might mean the measurement of PWM in some studies and EMW in some others. He maintains that PWM and EWM, relate to different functions of the system and impact upon different aspects of language learning and processing; that is to say, PWM generally exerts an impact on developmental and acquisitional aspects of learning (e.g. vocabulary, formulaic language or grammar) while EWM effects are mainly (Mitchell, Jarvis, O'Malley, & Konstantinova, 2015) implicated in processing and performance-related aspects of language (e.g. real-time performance). Therefore, studies should clarify what their tests aim to measure with a consideration of the language aspect under investigation; i.e. simple span tasks to measure PWM and complex span tasks to measure EWM. Related to this, Wen (2015) further argues that studies could limit their claims about WMC to what simple and complex span tasks measure.

Another issue is that WMC positively relates with L2 proficiency or L2 knowledge (i.e. knowledge residing in the LTM) (Mitchell et al., 2015). This suggests that previous knowledge of language may be a confounding factor in WM studies; for example, in a reading span task, learners might rely on their previous knowledge when they recall the target words or process the sentences. Finally, Gathercole (2007) notes that WM cannot be reliably estimated from a single task because 'memory typically involves access to and support from knowledge that is not stored in the working memory system' (p. 157).

Two implications emerging from the discussion above could be: 1) there seems to be no single pure measure of WMC, however, some tasks such as span tasks) may be effective tools to tap to into particular components (i.e. PWM and EWM) and 2) when selecting tasks regardless of the type (whether simple or complex span), reliance on language should be eliminated or minimized as much as possible. Also, Wen (2015) posits that at initial stages of L2 proficiency, PWM may play a larger role in language learning, while at advanced levels, PWM effects become less noticeable (i.e. becoming more L1-like) and EWM effects begin to surface more. Therefore, a third implication that would follow is that from a developmental perspective, it would be advisable to employ simple span tasks, which measure PWM, at lower levels, and to utilise complex span tasks, which measure EWM, as an additional measure for learners at more advanced levels (Juffs, 2006).

In the following section, I will discuss briefly how WM is related to the L2 oral performance, and L2 fluency more specifically, and what roles are assigned to the WM system in Levelt's speech production model.

2.5.2.3. Working memory capacity and L2 oral performance

WMC roles in L2 oral performance have been implicated in several L2 studies in relation to CALF dimensions (Mota, 2003), task variables such as different conditions or task features (Kormos & Trebits, 2011) or language proficiency (e.g. Kormos & Safar, 2008; Mitchell et al., 2015), development of certain speech aspects such as grammatical abilities (O'Brien et al., 2006) or L2 oral fluency specifically (Mizera, 2006). Regarding the fluency aspect examined in these studies, overall it is suggested that correlations exist between WMC and fluency with varying degrees; e.g. significant moderate correlations for speed fluency in Mota (2003), positive fairly modest correlation for speech rate in Gilabert and Munoz, 2010, weak positive correlations between for a range of fluency measures in Mizera (2006) and negative correlations for number of pauses in Georgiandou and Roehr-Brackin (2017). It was also implicated that LP was a mediating factor in the correlations between WMC and fluency performance (e.g. Georgiandou & Roehr-Brackin, 2017).

At this point, it is important to provide a clear explanation, from a theoretical point of view, on in what ways individual differences in WMC can impact L2 fluency. Although Baddeley and Hitch (1974)'s model does not specify this, one hypothesis is that those with higher WMC could be at an advantage in L2 speech in that they can manage the processing of the input and the output arguably better because higher capacity would allow them to free up more attentional resources and direct them elsewhere in the processing. A less-direct effect of this could be that higher capacity increases the extent that the declarative knowledge is proceduralized or automatized (Segalowitz, 2010) (e.g. more resources could be dedicated to the rules or the strategies), and higher degree of proceduralization and automaticity subsequently mean that there will be less reliance on WM itself in speech thanks to the increased speed and efficiency in lexical retrieval (Mota, 2003). In both hypotheses, limited resources in WM would be managed more efficiently, and speech processing would result in less disrupted and more fluent overt speech, further suggesting that learners with higher WMC would perform better on tasks examining speech fluency (Daneman, 1991).

As already mentioned, the existing body of research examining the role of individual differences in WMC on L2 speech fluency overall point to a link between the two, however, most of them examined fluency in a rather restricted perspective; i.e. with regard to speed fluency aspect only (e.g. Gilabert & Munoz, 2010; Mora, 2003; Kormos & Trebits, 2011). Therefore, other fluency aspects (i.e. breakdown and repair) seems to have largely been overlooked although some exceptions exist (e.g. Mizera, 2006, Georgiandou & Roehr-Brackin, 2017). More importantly, no study, to the best of my knowledge, has investigated whether individual variations in WMC mediate the relationship between L1 and L2 fluency behaviour, which is a gap Study 2 aims to help fill. Given that L2 learners face additional challenges in the speech production process, which require them to manage their attentional resources in the best way, it can be assumed that being at an advantage in WMC will support the process for L2 learners at especially lower levels. Therefore, it can be hypothesized that individual variations in WMC could mediate the relationship between L1 and L2 fluency behaviour by aiding the performance especially at lower levels.

2.5.3.Length of residence

L2 researchers have attempted to examine the influence of a number of additional factors on L2 speech, including motivational, emotional or social factors (Saito & Brajot, 2013). The role of the amount of L2 experience has been no exception. Often examined in study-abroad contexts, learners' linguistic experience in a TL community has received substantial interest from L2 several researchers (e.g. Mora & Valls-Ferrer, 2012; Saito, 2013, 2015). In these contexts, L2 experience has usually been defined in terms of their length of residence (LoR) abroad. The assumption is that lack of L2 input is considered as a constraint to learners' development of new linguistic competences in SLA, and L2 experience in this sense presents learners many more opportunities for language input. For example, learners could be engaged in authentic communication situations with native and non-native speakers and could become more exposed to different aspects of language, most probably on a daily basis.

It is widely believed that study-abroad experience is highly beneficial for developing L2 abilities, though not leading to equal improvements in all areas (Kinginger, 2011; Llanes, 2011). One of the areas that seem to benefit most from study-abroad has been the development of speaking abilities. Indeed, most research, in this regard, have reported linguistic gains in different aspects of speaking, such as phonological/pronunciation abilities (Díaz-Campos, 2004) oral proficiency (Hernandez, 2010), communicative competence aspects (Masuda, 2011), vocabulary (Dewey, 2008) and oral fluency (Du, 2013; Freed et al., 2004). Among these, speech fluency development has been one area where the largest gains have been observed consistently (Du, 2013; Huensch & Tracy-Ventura, 2017; Kim et al., 2015; Mora & Valls-Ferrer, 2012; Towell et al., 1996).

It is suggested that LoR could be considered as ‘a rough proxy for L2 experience’ (Saito, Ilkan, Magne, Tran and Suzuki, 2018, p. 597). Some researchers argued, however, that it does not necessarily mean that learners spend that amount of time interacting in the TL. For instance, Flege and Liu (2001) note that during residency abroad, some learners could opt to use their L1 more than their L2 as their main language for communication. Still, research evidence suggests that as long as there is some interaction between learners and native and non-native speakers during the time they reside abroad, learners can develop their L2 learning as a result of their increased LoR (Derwing & Munro, 2013).

To this end, studies have adopted several approaches to examining experience effects on L2 fluency development. Some researchers compared learners’ fluency development in the context of study-abroad and domestic/at-home instruction and found significantly more fluency benefits for those learners who study abroad (Serrano, Llanes, & Tragant, 2011). Yet, some others examined the fluency development for the same group of learners before and after they resided abroad (e.g. Di Silvio et al., 2016; Huencsh & Tracy-Ventura, 2017; Kim et al., 2015; Towell et al., 1996); their results overall pointed to significant fluency gains after residency abroad. While the existing research provide ample evidence about LoR effects in favour of L2 fluency development, what it does not clearly show is whether all aspect of fluency benefit from LoR equally. The overall picture suggests that learners benefit from L2 experience particularly in speed fluency aspect (i.e. speech rate) (e.g. Di Silvio et al., 2016; Du 2013; Huensch & Tracy-Ventura, 2017; Mora & Valls-Ferrer, 2012; Towell et al., 1996). Gains in breakdown fluency has also been reported (Trofimovich & Baker, 2006) though not consistently; in Kim et al. (2015) for example, learners produced more filled and silent pauses after residency abroad. As for the repair fluency, the picture seems less clear; some studies reported no gains in repair fluency (e.g. Di Silvio et al., 2016; Huensch & Tracy-Ventura, 2017). In fact, in early studies, it was suggested that repair behaviour may be a more evident feature of advanced learners' speech as these learners have a tendency to reformulate their speech more often and produce more false starts (e.g. Freed, 1995; Lennon, 1990). This was attributed to the *increased awareness* of mistakes in their speech, which comes out along with *increased proficiency.*

Although it is not well-established to what extent each fluency aspect benefit from L2 experience, speech fluency has been recognized as one area that benefits most from L2 experience. In other words, the more L2 exposure learners have, the more fluent their speech becomes. Given that Study 2 is situated in a study-abroad context, participants' L2 fluency development is likely to be affected by their L2 experience in the TL community, which is operationalized as LoR in the study. Therefore, it is plausible to hypothesize that LoR effects may also be at stake in the possible L1-L2 fluency associations, and as such it is deemed necessary to examine whether LoR, as another learner variable, mediates these relationships.

Apart from learner factors (i.e. LP levels, WMC, and LoR in this context) which have potential impacts on L1-L2 fluency links, there are also external factors. One of these relates to task variables which I will review as a final factor.

2.6. External factors

2.6.1. Task structure

A task can be defined as ‘an activity in which ‘meaning is primary, there is a communication problem to be solved, there is relationship to real-world activities; completion has a priority; and task assessment is based on its outcome’ (Skehan, 1998, p. 96). Studies within the CALF field have commonly utilised tasks as a means of eliciting speech samples. To this end, task performance has been analysed and assessed in terms of performance descriptors (i.e. CALF) to understand gains in L2 development or LP proficiency (Housen & Kuiken, 2009). Within a psycholinguistic approach to task-related practices, this is reasonable since L2 learners’ task performance can give us useful insights into the demands of a given task and the cognitive processes learners are engaged, and hence, could ultimately inform L2 practices, including theories, teaching and assessment.

Yet, there is ample research evidence suggesting that ‘tasks are not neutral vehicles’ (Tavakoli & Foster, 2011, p. 442) and a variety of factors can have consequences in task performance, including 1) task modality, i.e. written versus spoken tasks (Kormos & Trebits, 2012), 2) the conditions under which they are performed, e.g. planning conditions (Ahmadian, Tavakoli, & Vahid Dastjerdi, 2015) or task repetition (Ahmadian, 2012a; Ahmadian & Tavakoli, 2011; Lambert, Kormos, & Minn, 2017) and 3) specific features of tasks such as task type (e.g. picture narratives or topic response) or task design features (e.g. storyline complexity/intentional reasoning) (Awwad, 2017; Awwad et al., 2017) or task structure (Ahmadian et al., 2015; Tavakoli & Foster, 2011; Tavakoli & Skehan, 2005). Such task variables pertain to task demands or how complex a task is (i.e. task complexity), which is defined by Robinson (2001) as ‘the result of attentional, memory and other information processing demands imposed by the structure of the task on the language learner’ (p. 29). From a theoretical viewpoint, one important argument related to this is that manipulation of the information in a given task (e.g. complexifying or putting time pressure) would impact on speech processing at different stages of Levelt’s model, by either easing the burden or making additional cognitive demands on the speaker

(Skehan, 2015). Thus, the degree of task complexity will determine the amount of processing required, and it will eventually impact how learners will perform.

The cognitive perspective taken in Study 1 is that a complex activity such as speech production will require learners to divide their attentional resources to different stages of the production process (e.g. conceptualization or formulation) and that external factors such as task features may exert additional demands on learners for the management of their attentional resources in this process. Study 1 is interested in the effects of *narrative task design* features on fluency performance in terms of *task structure* only. It focuses on task structure among other possible task characteristics because task structure as a design feature has been consistently shown to affect fluency across different studies (e.g. Foster & Tavakoli, 2009; Tavakoli & Foster, 2011; Tavakoli & Skehan, 2005; Skehan & Foster, 1999) while other variables (except planning time which is a performance condition) (Ahmadian et al., 2015) did not demonstrate a consistent effect on fluency performance. Further information about narrative tasks and how task structure is examined in Study 1 will be provided in the methodology chapter (Section 3.5.3). Still, I should briefly note here that task structure is concerned with the underlying macrostructure of a task. For example, in an oral narrative task, it relates to how closely the picture prompts in the task are connected to each other. While a tight structure would mean the prompts have a sequenced order (with a clear beginning, middle and end), a loose structure means the order of the prompts could be changed or rearranged without comprising the story (De Jong & Vercellotti, 2016; Tavakoli & Foster, 2011).

Task structure as an external factor is investigated in Study 1 only in terms of whether performance on two tasks with different degrees of inherent structure (i.e. structured and unstructured tasks) mediates the relationship between L1 and L2 fluency behaviour. It is well-recognized that not only individual factors (i.e. LP, WMC and LoR) but external factors too have a bearing on L2 fluency performance, and as such, an investigation of the effects of task structure on L1 and L2 fluency seems highly relevant as well as necessary. Presence or absence of structure is assumed to affect cognitive fluency (i.e.by promoting or hindering it). This means performance in a structured task would likely to lead to greater utterance fluency. One can therefore hypothesize that performance on a structured task would also be associated with stronger links between L1 and L2 fluency, which is that Study 1 aims to investigate.

2.7. Summary of the chapter

In this chapter, I have presented an overview of the literature on second language oral performance and discussed the underlying rationale for research in Study 1 and Study 2. Starting with a description of popular models of L1 and L2 speech production, I discussed in what aspects L2 speech differs from L1 speech. Then, I moved on to describe L2 oral performance within CALF framework bringing forth its two important aspects (i.e. fluency and lexical complexity), while explaining that Study 1 and Study 2 would be centred mainly on the examination of fluency and lexical complexity formed one part of Study 2. Allocating a larger space for fluency, I have discussed several key issues relating to the definitions, the operationalizations and the measurements of these two constructs. I also highlighted that fluency and lexical complexity have been shown to be two of the best and most reliable predictors of L2 proficiency, and further discussed how they could be related to each other from a theoretical perspective.

I indicated that I would adopt Segalowitz (2010)'s categorization of fluency, namely cognitive fluency, utterance fluency and perceived fluency. I described each of these and provided a more detailed discussion on utterance fluency, which would be the main focus of Study 1 and 2, together with its different aspects (i.e. breakdown, repair and speed) and fluency measures for each. I also presented an in-depth discussion of the lexical complexity construct and showed how definitions and methodological approaches to the investigation of this construct in studies vary greatly and present challenges for L2 researchers. I situated lexical complexity within the CALF context and reviewed its three dimensions: LD, LS and lexical density. I pointed out that lexical complexity would be examined in Study 2 only in terms of LD.

Next, I have presented detailed discussions on three individual learner factors, i.e. LP, WMC and LoR, and one external factor, i.e. task structure, with several sections dedicated to each. I indicated that each of these factors influence L2 fluency performance (although in different ways), and therefore each could potentially mediate L1 and L2 fluency relationships. I also explained that task structure would be examined in Study 1 using two narrative tasks (i.e. structured and unstructured), while the mediating effects of WMC and LoR would be explored in Study 2. LP would be examined in both studies, with different targeted learners. In each study, LP would be viewed in terms of declarative vs procedural knowledge distinction and be measured through two standardized tests; i.e. OPT to represent declarative knowledge and EIT to represent procedural knowledge. In the next chapter, I will present and discuss the methodology of Study 1.

CHAPTER 3. STUDY 1 METHODOLOGY

3.1. Introduction

In this chapter, I present a discussion of the methodological approaches used in Study 1 in twelve sections. Beginning with the aims of the study, I provide the research questions and the corresponding hypotheses that guided this research. Then, I move on to describe the research design, variables of the study and the research instruments employed with justifications provided for using each tool. I also present a detailed discussion of pre-pilot and pilot studies and explain in-depth the procedures followed in Study 1 including the setting of the study, participants, and data collection stages. These are followed by how data were analysed, which covers data transcriptions, fluency measures adopted and data coding, and how interrater reliability for each different stages of analysis was ensured. The final section provides a summary of the chapter.

3.2. Aims of the study

Study 1 set out to explore whether there was a relationship between L1 fluency (Turkish) and L2 (English) fluency behaviours for learners of low levels (A2, B1 and B2 at CEFR) in an EFL context. As discussed in the previous chapter, the motivation for the study comes from the existing research evidence which suggests that L1 fluency behaviour is, at least to some extent, mirrored in L2 fluency behaviour. Study 1 attempts to expand our knowledge of the link between L1 and L2 fluency performance in several directions. First, it aims to explore L1-L2 links between a typologically different language pair (i.e. L1 Turkish and L2 English) as opposed to most studies in this field which investigated fluency in typologically closer languages (De Jong et al., 2015; Derwing et al., 2009). Turkish is a highly agglutinative language, where words can take several affixes to create a new word or meaning; as such, grammatical functions or different meanings could be expressed through these affixes (mostly suffixes). For example, a single word in Turkish such as 'konuşuyorlar', with 'talk' (konus-) as the root word, can be expressed as a sentence in English (i.e. ‘they are talking’). This implies that the number of syllables for any given two utterances produced in the two languages is likely to be different even though intended meanings are the same. Although this does not seem to have any implications for the calculation of fluency measures which are based on the number of syllables produced (e.g. speech rate) in a given amount of time (e.g. a minute), this difference is likely to affect the rate at which syllables are produced. Given that Turkish speakers are likely to produce a word with a lot of syllables to represent a whole sentence, it may well be assumed that this could have an impact on speed with less opportunities for pausing.

Secondly, Study 1 seeks to understand whether variations in L2 proficiency mediates the strength of the relationship between L1 and L2 fluency behaviours. Research in SLA has indicated that LP, as an individual learner variable, is directly related to language development and acquisition (Hulstijn, 2012, 2015). It has been recognized that the degree of automaticity in L1 and L2 speech is different although similar processes essentially take place in both (Kormos, 2006; Skehan, 2015). Parallel and automatic processing mechanisms in L1 speech makes the speech fluent without or less conscious attention from the speaker. However, L2 speech production is not yet automatized and processing requires more conscious attention from the speaker. This is likely to result in signs of disfluencies in the overt speech such as slower speech or frequent pauses, especially for learners at lower proficiency levels due to incomplete linguistic knowledge (Kormos, 2006). As learners improve in LP, processing becomes less demanding and more automatized. Therefore, individual variations in LP are predicted to have an impact on the relationship L1-L2 fluency behaviours by overshadowing the performance at lower levels.

In addition, Study 1 aims to examine LP from a broader perspective by using two LP measures (the OPT and the EIT) each of which is assumed to test one kind of linguistic knowledge; i.e. the former declarative knowledge and procedural knowledge. As was discussed earlier in 2.5.1.1., this research perceives LP as consisting of these two underlying constructs. Yet, most fluency research has examined either of the two knowledge types only. This is a methodological shortcoming of the research in this field because certain aspects of fluency (e.g. breakdown or speed) have been suggested to be linked with different types of linguistic knowledge (e.g. Tavakoli et al., 2019). Also, compared to most fluency studies that have examined only one or two proficiency levels (e.g. De Jong et al., 2015; Derwing et al., 2009; Huensch & Tracy-Ventura, 2016), Study 1 is one of the few studies in which participants belong to three different proficiency levels (i.e. A2, B1 and B2). Clearly, using two measures of LP to test both knowledge types and examining three LP levels would allow a broader picture of learners' LP and a more valid interpretation of the mediating role LP plays in the strength of the relationship between L1 and L2 fluency behaviour.

Finally, the present research also aims to look at whether task structure as an external factor is a mediating factor in the strength of the relationship between L1 and L2 fluency behaviours. The rationale for incorporating task structure as another independent variable comes from previous research (Foster & Tavakoli, 2009) suggesting that task structure, i.e. loose versus tight, affects L1 and L2 fluency differently. When a task consists of very obviously sequenced episodes (i.e. tight structure), fewer processing demands are made in the conceptualization stage (and possibly in the formulator stage as well) and this leads to increased fluency. Therefore, it is hypothesized that task structure would mediate also the strength of the relationship between L1 and L2 fluency behaviours.

3.3. Research questions and hypotheses

Study 1 has been guided by the following research questions and hypotheses:

RQ1) To what extent are L1 Turkish and L2 English fluency behaviours related among lower-proficiency L2 learners in an EFL context?

H1. L1 and L2 fluency behaviours will be related to each other at least to some extent for all fluency aspects (i.e. breakdown, repair and speed).

RQ1a) Does level of L2 proficiency, measured through the OPT and the EIT, mediate the relationship between L1 and L2 fluency behaviours among lower-proficiency L2 learners in an EFL context?

H1a. L2 proficiency will mediate the relationship between L1 and L2 fluency behaviours, especially at lower levels.

RQ1b) Does task structure mediate the relationship between L1 and L2 fluency behaviours among lower-proficiency L2 learners in an EFL context?

H1b. Task structure will mediate the relationship between L1 and L2 fluency behaviours by affecting cognitive fluency.

RQ2) To what extent can L2 fluency behaviour be predicted from both L1 fluency behaviour and language proficiency scores among lower-proficiency learners in an EFL context?

H2. L2 Fluency measures for all aspects of fluency (i.e. breakdown, repair and speed) will be predicted from their equivalents in L1 and each L2 proficiency score, at least to some extent.

3.4. Design

Study 1 adopts a *quantitative* approach in which 'data collection procedures mainly result in numerical data' and statistical methods are primarily used to analyse the data (Dörnyei, 2007). Examples of quantitative data involve the data obtained using survey questionnaires and language tests. In quantitative research, researchers usually study the relationship between variables and data is quantified or converted into statistics (Cohen, Manion, & Morrison, 2011, 2018; Lowie & Seton, 2012; Mackey & Gass, 2011) upon which judgements can be made about degrees of significance. Since the current study investigates the relationship between L1 and L2 fluency behaviour through tests or measures which require the use of numerical data such as scoring proficiency tests or calculating the fluency measures (e.g. number of pauses or length of pauses), this investigation falls within quantitative research.

Since the current study had multiple independent variables (i.e. factors), a mixed within-between factorial design was employed to investigate the effects of different variables on participants' fluency performance in L1 and L2. A factorial design typically looks at how one manipulated variable (or a combination of more than one) affects a dependent variable; yet they can also include non-manipulated independent variables as well such as in correlational studies. The study had two independent variables, i.e. LP and task structure, both of which are between-subject variables; with LP described with three levels (i.e. A2, B1 and B2) and task structure described with two levels (i.e. tight and loose structure). Participants' fluency performances in the two languages are represented through fluency measures and these are within-subject variables of the study. It is worth noting that task structure was used as a between-subject variable (rather than a within-subject variable) for practical purposes; that is to say, although as a within-subject variable task structure would have provided valuable insights into the effect of task structure on L1-L2 fluency links, it could have led to losing some of the participants as they would have been asked to do more tasks. The design of Study 1 and the variables are presented in Table 3.1 below.

Table 3.1. Study design and the variables of Study 1.

Study design	Independent variables	Dependent variables
Mixed factorial design within-between participants (N=42)	▪ L2 proficiency with three levels: - A2, B1 and B2 levels at CEFR ▪ Task structure with two levels: - Tight and loose structure	▪ L1 speech fluency performance (represented through fluency measures) ▪ L2 speech fluency performance (represented through fluency measures)

3.5. Instruments

3.5.1. Background questionnaire

A short demographic questionnaire was designed with the purpose of controlling for learner individual variables such as age, educational background or exposure to the language (Appendix A). The questionnaire consisted of two sections. The first section enquired about participants' demographic information (e.g. age, gender, level of education, etc.) whereas the second section aimed to elicit information about participants language profile (e.g. the number of hours they contact with English outside the class.)

Previous research has indicated that L2 experience abroad or the use of language for professional purposes in everyday life would help develop fluency faster (Mora & Valls-Ferrer, 2012; Tavakoli, Campbell, & McCormack, 2016). In order to address the research questions of the present study, a group whose L2 fluency was not influenced by such factors was needed. It was deemed important to ensure that the sample group was homogenous in such aspects and there was not large variation amongst them that could affect the results. With the intention of controlling for these, the questionnaire was given to the participants before the data collection, which made it possible for the researcher to determine as early as possible whether the participants were eligible for the study or not. As such, any participants who had lived/studied abroad before or used language for professional purposes were excluded from the study.

3.5.2. Proficiency tests

In SLA research, proficiency has been measured rather variably, ranging from the use of a single test such as a vocabulary or cloze-test tests to the use of shorter versions of standardized tests (e.g. Oxford Placement Test or Cambridge University Test) or internationally validated four-skill tests (e.g. IELTS). As was mentioned in Section 2.5.1., following Hulstjin's (2015) definition, LP assessment in the current study was based on the two dimensions of 'knowledge of language' or *declarative knowledge* and 'the ability to use the language' or *procedural knowledge.* As such, the LP assessment tools were selected to represent each knowledge types (although neither of them could be assumed to test *exclusively* one knowledge type): standardized tests of OPT (Alan, 2004), which was more likely to elicit declarative knowledge and EIT (Ortega, Iwashita, Norris, & Rabie, 2002) to elicit procedural knowledge. Measuring LP from a broader perspective was intended to provide us with a more accurate assessment of participants' linguistic knowledge (Elder & R. Ellis, 2009), and a more valid interpretation of the role LP plays in the strength of the relationship between L1 and L2 fluency behaviours.

At this point, as was mentioned earlier in Section 2.5.1, I should remind the reader that no proficiency test was used to measure proficiency in Turkish because all the participants of the study were assumed to have similar educational background as were all studying in their bachelor's degree programs in Turkey at the time of the testing.

3.5.2.1. Oxford placement test

The OPT, a widely used standardized test, was originally designed by Dave Alan in 1985. It was later developed and made commercially available by Cambridge ESOL and Oxford University Press. The OPT involves two parts, i.e. listening and grammar sections each consisting of 100 test items; however, for reasons of practicality, only the grammar section (a pen-and-pencil multiple-choice test), was employed in the study (Appendix B). The test consisted of a range of grammatical and lexical items, presenting test takers three alternatives for each question (e.g. “He didn’t stop *playing/to play/play* for Santos”) and asking them to select the correct answer.

The test widely used in other studies of this kind (Awwad, 2017; Awwad & Tavakoli, 2019; Declerck & Kormos, 2012; Gilabert & Munoz, 2010; Malicka & Levkina, 2012) is considered as a valid and reliable tool to assign participants to different proficiency groups. It can also be considered as a practical tool both in terms of the administration (e.g. no special equipment is required), and scoring. However, although the use of the OPT has been validated across several studies, it was deemed necessary to complement the results of this test with another tool for two main reasons: 1) a sole reliance on a grammar test to measure L2 proficiency was considered as a potential limitation (Awwad, 2017; Awwad & Tavakoli, 2019) since the proficiency construct was likely to be underrepresented, and 2) given that the focus of the study was on the investigation of oral performance and the OPT was assumed to measure declarative knowledge[2] mostly,

---

[2] Although I fully acknowledge that a grammar test can partly tap into procedural knowledge as well, for the purposes of the current study, it was primarily used as a means of activating declarative knowledge.

another tool which required some language production and would tap into procedural knowledge as well was required. Therefore, the results of the OPT were complemented with an oral proficiency assessment tool, i.e. the EIT.

3.5.2.2. Elicited imitation task

The EIT is a measure of procedural knowledge test (Gaillard & Tremblay, 2016; Ortega et al., 2002; Wu & Ortega, 2013; Yan, Maeda, Lv, & Ginther, 2016), which has been increasingly used in SLA studies as a valid and reliable proficiency test (Erlam, 2006; Kim, Tracy-Ventura, & Jung, 2016). In the test, participants are asked to listen to a set of sentences in varying length and complexity (measured by syllables) and to repeat these as exactly as possible. Since 1970s, this tool has been used to test a range of phenomena such as implicit knowledge, morphology, listening comprehension or L2 proficiency (Verhagen, 2011) as well as to test performance on a range of specific structures such as syntactical, lexical, phonological (Akakura, 2012; West, 2012) While an increasing number of studies employ EITs to tap into measure the *implicit linguistic knowledge*, i.e. "mental representation of linguistic knowledge and automaticity in language processing" (Yan et al., 2016, p. 11), given that the current research makes a distinction between declarative versus procedural knowledge types, rather than explicit versus implicit knowledge types, EIT has been adopted here as a measure of *procedural linguistic knowledge* (also discussed in Section 2.5.1.1).

The theoretical rationale for adopting EIT as a measure of procedural knowledge comes from the assumption that in the test, test takers are required to process the stimuli presented to them and reconstruct the meaning using their interlanguage system, i.e. their resources already available (R. Ellis, 2005). In other words, in order for the test-takers to repeat the stimuli accurately, the target structures must be available in their interlanguage system (i.e. declarative knowledge stores), and must be put into use (i.e. procedural knowledge) (DeKeyser, 2017). Additionally, since the EIT is dependent on oral skills in that test-takers are required to process and produce language under time constraints, it is arguably more suited to measuring procedural oral language ability; i.e. both implicit and automatized explicit knowledge types[3] (Suzuki & DeKeyser, 2019).

Some researchers have expressed their concerns about the validity of EITs. For example, it has been argued whether test takers of EIT are engaged in language comprehension or simply rote repetition, i.e. repetition without comprehension (Vinther, 2002); yet, the main assumption underlying this kind of testing is that test-takers must first *comprehend* and decode structures displayed in the stimuli, so that they can recall and reconstruct them. Only then will they be able to accurately repeat the stimuli (e.g. Rebuschat & Mackey, 2012). Additionally, when the length of sentences exceeds working memory capacity, it is suggested that people will not be able to parrot them. L2 research has indeed provided a great amount of evidence in favour of this *reconstructive* nature of EITs (e.g. Erlam, 2006; Gaillard & Tremblay, 2016; Ortega et al., 2002). Further related to the validity of EITs, a recent meta-analysis by Yan et al. (2016) examined 76 studies which employed the EIT, and the findings overall pointed at the EIT's potential as a valid and reliable proficiency assessment tool. Although how the knowledge is processed cannot be observed directly, EITs have also commonly been used to distinguish LP levels (e.g. Erlam, 2006; Gaillard & Tremblay, 2016, Wu & Ortega, 2013). Since higher proficiency speakers have higher levels of automaticity in their speech production process and are capable of accessing their linguistic knowledge due to a wider range of lexical and syntactical structures which have been internalized in their system, these speakers are also assumed to be more capable of repeating longer and more complex structures and be less dependent on rote-repetition (Yan et al., 2016). Speakers of lower proficiency, in contrast, will more likely depend on rote repetition since their access to their linguistic sources will be limited. It is also worth mentioning that several studies showed that EIT scores

[3] Since the current study does not distinguish between implicit knowledge and automatized explicit knowledge types, the use of EIT seems well justified for measuring procedural knowledge of both kinds.

correlate with specific aspects of L2 proficiency; e.g. oral performance in Wu and Ortega (2013) or lexical complexity in Tracy-Ventura, McManus, Norris and Ortega (2014).

As already mentioned, EIT was employed in the current study as a psycholinguistic proficiency assessment tool, i.e. focusing on language processing and automaticity (Van Moere, 2012) which would complement the limitations of the OPT and help measure LP from a broader perspective. The test was adopted from Ortega et al. (2002). The test is comprised of a combination of grammatical features, syntax and vocabulary; there was a total of 30 sentences in English, which were of varying lengths (7-19 syllables) and was ordered from the fewest syllables to the highest. Research has shown that using varying sentence lengths are more sensitive in discriminating different proficiency levels (Serafini, 2013) as learners from higher levels are more capable of repetition of longer sentences. The task also involved a practice session at the beginning of the test, using 5 additional sentences in Turkish, to ensure that task procedures were well understood and followed.

3.5.3.Oral narrative tasks

Narrative tasks are defined as ‘stories based on a sequenced set of picture prompts, which are given to participants in order to elicit language performance’ (Tavakoli & Skehan, 2005, p. 248-9). The cognitive demands of talking in time in an oral narrative and the degree of control in the linguistic units it elicits have made oral narratives a popular research instrument in studies examining L2 speech (Albert & Kormos, 2011; Kormos & Trebits, 2012; Tavakoli, 2009). However, specific features of narratives affect the language production in different ways. As was discussed in Section 2.6.1., the narrative feature that the current study is interested in is *task structure*, i.e. how connected the picture prompts are to each other (De Jong & Vercellotti, 2016). For this purpose, two oral narrative tasks, which differed in terms of the degree of structure, were used; a tightly structured task and a loosely structured task. A narrative with a ‘tight’ structure had a clear time sequence of prompts, i.e. a clear beginning, middle and end to the story, while a narrative with a ‘loose’ structure was comprised of picture prompts or ‘events which can be reordered without comprising the story’ (Tavakoli & Foster, 2011; p. 447). The assumption underlying the distinction of tight vs loose structure was that when a task is not comprised of obviously sequenced episodes (i.e. loose structure), the performer is required to create meaningful links between the prompts, and this adds more processing and attentional demands on speech process (largely on processes taking place in conceptualization and formulation stages). A task with a tight structure, on the other hand, is assumed to ease the processing burden, allowing more attentional resources to be devoted to fluency (and accuracy) (Tavakoli & Foster, 2011). Studies which tested these assumptions (Skehan & Foster, 1999) reported effects of task structure on fluency. Interestingly, task structure seemed to help non-native speakers' performance (Tavakoli & Foster, 2011) while no significant effects were observed on native speakers' performance (Foster & Tavakoli, 2009).

Yet, De Jong and Vercellotti (2016) claim that picture prompts used in narrative tasks may have confounded the findings in this line of research. The authors argue that although researchers most commonly pre-determine the equivalence of picture prompts in parallel tasks and assuming that the tasks elicit similar speech, the prompts in each task might differ in additional dimensions of task design. They set out to investigate whether or to what extent features of story prompts influence performance in narrative tasks, and their results indicated that different prompts influenced the performance in fluency and lexical aspects. Based on several possible explanations for this, which they discuss in their study (p. 400), the authors recommend L2 researchers to consider a number of elements (i.e. main characters and an props), in addition to task structure and storyline complexity, when selecting and/or designing their tasks.

Following from De Jong and Vercellotti's (2016) recommendations for careful selection of "features that constitute task complexity" (p. 387), a number of criteria were considered when selecting the narratives used in Study 1. These included the number of characters and props (De Jong & Vercellotti, 2016), amount of contextual support (Révész, 2014), similar storyline complexity (Tavakoli & Foster, 2011), and similar amount of intentional reasoning (Awwad, et al., 2017). In addition to these, some of the other criteria considered included 1) the clarity of the pictures and the story, 2) a reasonable length, 3) appropriateness for task-performers in terms of the vocabulary and the linguistic structures required to narrate the stories.

Although the initial aim was to select one task and adapt its structure in terms of the elements mentioned above in order to create a second one, a careful examination of a large number of narrative tasks in EFL resources (e.g. course books) and in online sources (e.g. videos or cartoon clips) proved this task to be almost impossible. Therefore, two different Henry tasks, which were assumed to be parallel to each other based on the above criteria, were selected and adapted for the purposes of the study (Appendix C). The tasks were named appropriately; i.e. Task A (the ice-skating boy) and Task B (the boy looking for his car). Task A was used as a structured task as it had a macrostructure with a visibly seen beginning, a middle and a reasonable end to the story and the order of the prompts was fixed (i.e. no rearrangement of the pictures was possible). Task B, on the other hand, was a less structured task; i.e. the sequence of the events was rather arbitrary, repeated in the middle and towards the end of the story. It was assumed that one performing on Task B could change the order of the events without compromising the main theme of the story.

It is also worth mentioning here that before being employed for the study, some adaptations were made on the original versions of both tasks. One picture prompt was removed from Task A to equal the number of prompts (i.e. 8 prompts) in each task while the sequence of the pictures was manipulated in Task B by replacing the last prompt with a new one. For the adaptation of Task B, professional help was sought from a colleague who was a lecturer in the Department of Graphic Design at a university in Turkey. The final versions of both narratives had the same number of prompts (8 in each) and characters (2 in each). Each prompt in the tasks was numbered appropriately and task instructions were provided at the beginning.

3.6. Pilot study

A pilot study helps the researcher to eliminate 'some of the inevitable problems of converting the design into reality' (Robson, 2002, p. 383). As such, before the commence of Study 1, a pre-pilot and a pilot study were designed to find out whether the design was in line with the aims of the study and whether any amendments were needed in terms of the materials, participants, procedures and measures used. Therefore, the pre-pilot and pilot were carried out as small replications of Study 1.

Six native speakers of Turkish were recruited: 5 females and 1 male. Although every effort was made to ensure that the participants had a similar profile (e.g. with regard to their age, first language background, educational level, etc.) to the target participants of the main study, for practical reasons both the pre-pilot and the pilot studies were carried out with Turkish post-graduate students in the UK. Their proficiency levels, measured by the OPT and the EIT, ranged between B1 and B2 and they were aged between 27 and 32. Three of them participated in the pre-piloting of the proficiency tests and the oral narratives, and the remaining three took part in the piloting of all materials.

In the pre-piloting, the participants first completed the proficiency tests in individual sessions; i.e. first taking the OPT and then the EIT. When taking the OPT, they were told to answer all questions and write their answers in the space indicated (Appendix B). In order to be able to estimate the average amount of time needed to complete the test in the main study, they were not limited by time. After they completed their test, they moved onto performing the EIT. They were given the task instructions (Appendix D) in Turkish and did a practice session in Turkish before starting the test. When it was ensured that the procedures of the test were well understood, they performed the EIT. Finally, the participants were given the oral narratives. They were assigned to two groups; with one group to perform in Task A and the other to perform in Task B in both languages. Their performances in the EIT and the narratives were recorded on a digital audio recording in order to detect any technical issues in advance, and there was a break of approximately 20 minutes between the end of the EIT and the start of the narratives.

The results of the pre-piloting indicated that the task instructions were clear, and the participants did not have any problems in understanding them. However, the participants reported that they became tired from performing all the tests/tasks on the same day. It appeared that since the first test, i.e. OPT, was quite a long test for the participants, they could not concentrate effectively on any other tests which they performed after that. This was an important issue emerging from the pre-piloting, and the decision was taken to spread the data collection onto two separate days in the piloting study.

Once the pre-piloting was completed, all the materials were further piloted with another group of participants (Table 3.2). This time, the data collection took part in two individual sessions on separate days within the same week. On the first day, the participants performed the proficiency tests, i.e. first the OPT and then the EIT, which were preceded by the short demographic questionnaire. Since the OPT was reported to be a long test, a 10-minute break was given between the end of this test and the start of the EIT. On a separate day, the participants performed the oral narratives in two groups; with first group (n = 2) performing Task A and the other (n = 1) Task B in both Turkish and English. The order of the starting language was counterbalanced, and their performances in the second session were audio-recorded again. Upon the completion of both sessions, the participants were asked to raise any issues or concerns regarding the clarity of the materials and the instructions. They reported that they did not have any problems in understanding them and the materials were clear and appropriate.

Table 3.2. Tools, number of the participants and the sessions in the pre-pilot and pilot study of Study 1

(N=6)	**Number of sessions**	**Tools piloted**
Pre-pilot study (n=3)	One individual session (with a 20 mins break in between)	• All materials (an approximate of 90 minutes including the instructions and the break time) • The OPT • The EIT • The oral narrative tasks - one in Turkish and the other in English - the order of the starting language counterbalanced
Pilot study (n=3)	Two individual sessions	<u>Session one</u> (an approximate time of 60 minutes including the instruction time): • The background questionnaire • The OPT • The EIT (after a 10 min break) <u>Session two</u> (an approximate time of 10 minutes including the instruction time :

	• Two oral narrative tasks ○ one in Turkish and the other in English ○ the order of the starting language counterbalanced

3.7. Ethical considerations

This research followed the Ethics Guidance of Reading University. In order to ensure that the study and the data collection procedures adhered to the University's Ethics Guidance, I submitted an application for the approval to the Ethics Committee of the School as well as the Committee of the university from which the participants were recruited. The application was reviewed, and the ethical approval was granted by the Committees of both institutions (Appendix E). During the data collection, the processes were put in place to ensure the confidentiality and the privacy of the data. The participants were informed about their rights to withdraw from the study at any time they wished to with an information sheet (Appendix F) as well as verbally by the researcher. They were also ensured that their names or any identifiers would not be mentioned, and the data obtained would be kept safe and used for only research purposes.

All participants, including the ones in the pre-pilot and pilot study, were given the information sheet which clearly explained the purpose of the research and participants' rights and detailed the data collection procedures. It was confirmed with the participants that the information sheet was read and understood. Also, informed consent was sought from each participant before starting the data collection (Appendix G).

3.8. Procedures

In this section, I explain the data collection procedures followed in Study 1. This covers the context of the study, information about the profile of the participants and how they were recruited and an account of the data collection phase detailing the steps followed. This section also includes the information about how the data were analysed, the issues emerged in this process and a discussion of the fluency measures that were selected for the analysis.

3.8.1. Setting and participants

The participants of Study 1 were recruited using convenience sampling, i.e. choosing individuals who are available and reachable at the time of the study (Cohen et al., 2018). They were undergraduate students, studying at various degrees at a state-university in Turkey where English is taught as an EFL at most schools and is used in limited contexts such as for communication purposes with visitors of the country or for business purposes. As is the case in most universities in Turkey in general, the language of the instruction at this university was Turkish. The participants were taking English classes either as a compulsory course in the first year of their degree programme or as part of their foundation classes (i.e. one-year compulsory education before the start of their degree programme).

As was already mentioned, prior to the study, the formal approval for the data collection was sought from the university authorities. With the help of colleagues in the Foreign Languages Department, a short demographic questionnaire was given to all participants to ensure the homogeneity of the sample. The inclusion requirements for participants were 1) studying at a bachelor degree (at least) in any subject in Turkey, 2) being a unilingual native speaker of Turkish and 3) having no or little (less than three months) study abroad/life experience. Based on the research evidence suggesting that previous study-abroad/life experience would help learners develop their fluency faster (e.g. Mora & Valls-Ferrer, 2012; Tavakoli, et al., 2016); only those participants whose fluency has not been influenced by such factors were included in the study. In addition, it was assumed that having substantial knowledge of two languages might have an influence on fluency performance. Therefore, using the information obtained through the questionnaire, anyone who was not monolingual speakers of Turkish (e.g. those who were brought up as bilingual speakers of Turkish and Kurdish) were removed from the study.

The data were initially collected from 44 participants, however, at a later stage two of them were excluded from the study since they did not meet task requirements (i.e. they did not complete all tasks). The data reported here come from 42 participants; 25 females and 17 males, all native speakers of Turkish. They were aged between 19 and 25 (with a mean of 21) and belonged with three different proficiency levels (i.e. A2, B1 and B2 according to CEFR levels) measured by means of the OPT and the EIT. They had been studying English for five months at the time of the study and had varying language L2 proficiency levels at the start of their programmes.

3.8.2.Data collection

Drawing from the results of the pilot study which suggested that a long session would have a potential damaging effect on participants' performance, the decision was taken to spread the data collection procedure onto two separate days within the same week. Therefore, the whole procedure took place in two stages and was arranged and ran by the researcher.

Phase 1: The first session was arranged for all participants. For practical reasons, they first took the OPT (Appendix B) in a classroom, which was booked with the help of the department administration. Prior to the test, the participants were reminded about the aims of the study and they were given an information sheet and a consent form to sign. Test instructions were provided in Turkish which included that 1) they had 60 minutes to complete the test, 2) they needed to fill in all the questions without leaving any unanswered question and 3) only one answer was acceptable for each question. At the end of the test, they were also asked about their contact information, so that individual meetings could be set up with the researcher for the second phase of the data collection.

On the same day, the participants were met individually by the researcher in a quiet room for them to perform the EIT. In the implementation of EITs, the procedures described in Gaillard and Tremblay (2016) were followed. The test started with a practice session which included listening and repeating five sentences in Turkish. After it was ensured that the test procedures were clear and well-understood by the participants, they started the actual test. The test instructions were explained to each participant in Turkish. These included that 1) they were going to listen to a number of sentences in English and hear a beep sound after each sentence, 2) they were required to repeat each sentence after the beep sound as accurately as possible, 3) they had only one attempt to repeat the sentence,

4) they were not allowed to take any notes during the task, and 5) if they could not understand a sentence, they needed to wait till the next one is presented.

The EIT began after the practice session. The sentences, followed by a beep sound, were presented to the participants one by one. There were two seconds between end of each sentence and the beep sound. This arrangement was based on the suggestion in previous studies (Gaillard & Tremblay, 2016; Ortega et al., 2002) that in this way test takers would process the stimuli rather than simply mimic them. When they could not repeat and remained silent for five sentences in a row, the researcher ended the task. The test took up to 9 minutes depending upon each participant's performance. It is worth noting that one participant wanted to repeat the stimuli right after listening to the sentences, without waiting for the beep sound as he told the sound was confusing him, and he was told that he needed to wait till the beep sound. Furthermore, not all participants were available to do this test on the same day. Therefore, a small number of them took the EIT in the second phase of the study, which is described below.

Phase 2: On a separate day, the participants were met again individually. They had been randomly divided into two equal groups; with one group to perform in the tightly structured task and the other to perform in the loosely structured task. The picture prompts had been colour-printed on an A5 paper, and these were shown to each participant. They were given 30 seconds of planning time to familiarize themselves with the story before they were asked to start (Skehan, 2015) and 90 seconds to retell the story. The task instructions and all the information about the purpose of the study were provided in Turkish. No further information about the tasks or vocabulary to be used were given.

Each participant was asked to narrate the story twice; once in English and once in Turkish. The participants were allowed to look at the picture prompts while narrating the stories, and their performances were audio-recorded. The starting language of the tasks was counter-balanced in order to control for any practice effect; i.e. half of the participants in each group started narrating the task in Turkish first and then moved on to narrate it in English, and the other half performed the tasks in the reverse order (i.e. first in English and then in Turkish). The details of this counterbalanced design are shown in Table 3.3 below.

Table 3.3. The groupings of the participants across tasks in Study 1

Tasks (N=42)	**Sub-groups**	
Task A. (The boy going ice-skating boy) n=21	**Group 1** (English first, Turkish second) n= 10	**Group 2** (Turkish first, English second) n=11
Task B. (The boy looking for his car) n= 21	**Group 1** (English first, Turkish second) n=11	**Group 2** (Turkish first, English second) n= 10

3.9. Data analysis

In this section, the data analysis procedures are described and discussed, including how the proficiency tests (i.e. the OPT and the EIT) were scored, the steps followed the data transcription and the data coding as well as the identification of pauses using PRAAT software.

3.9.1. Marking of the proficiency tests

Both the OPT and the EIT tests were marked manually. The OPT comprised of 100 questions in total, and for each correct answer 1 point was awarded. The participants had already been reminded to complete all the items not leaving any unanswered questions. Although rather rarely observed, there were still some items left unanswered on the participants' test sheets, possibly due to a transfer effect from the examination system in Turkey. In most high-stake national exams (e.g. ALES), test takers are repeatedly reminded not to answer questions when they do not know or are not sure about the correct answer because they may be awarded partial credit for any unanswered question or may be penalized for incorrect answers (i.e. usually by removing a quarter point from test taker's total score) in order to discourage guessing. Given that an incorrect answer earns nothing in this test, any unanswered questions were considered as incorrect items, and no point was awarded for these. The maximum score that one could gain from the OPT was 100. In order to see whether all the items on the OPT could reliably measure the same construct (i.e. language proficiency), Cronbach's alpha coefficient was calculated based on the participants' scores and a reliability coefficient of .89 was obtained. This high score indicated the test was reliable.

The marking of the EIT test was also done manually based on a 5-point scale (0-4), which was adopted from Ortega et al. (2002) but was also utilised in several other studies (e.g.Tracy-Ventura et al., 2014; Wu & Ortega, 2013). The scale (Table 3.4) included four levels: 4 points were given for perfect repetition of the stimulus, 3 points for accurate repetition which could include minor changes in the structure but no change in the meaning, 2 points for stimulus which included changes in the meaning or the structure, 1 point for half of the stimulus repeated, and 0 point was awarded for no production or less than half of the sentence. The test takers' productions were rated one participant at a time, and the maximum score one participant could receive from the EIT was 120. When Cronbach's alpha was calculated for the EIT test, a coefficient score of .94 was achieved, showing that the test was reliable.

At this point, it is worth noting that most participants at lower-proficiency levels could not repeat the sentences after listening to the half of the sentences. Those at higher-proficiency levels, on the other hand, were able to continue performing till the end of the test and complete it successfully. This is likely because the participants at lower levels they were not able understand and process the sentences due to their incomplete linguistic knowledge (Kormos, 2006), and thus they did not want to engage in the task further. In such cases, the sentences which were not repeated at all were marked as zero.

Table 3.4. Marking procedures of the proficiency tests.

	Total score	Marking procedures	
The OPT (100 questions)	100 (later converted into 120)	▪ 1 point was awarded for each correct answer	
The EIT (30 sentences)	120	▪ Scored based on a 5-point scale (Ortega et al., 2002)	4 points= perfect repetition
			3 points= accurate repetition with minor structural changes and no change in the meaning
			2 points= repetition with changes in the meaning or structure
			1 point= repetition of half or less of the stimulus
			0 point= no or unintelligible repetition
		▪ An excel file was created with 30 sentences ▪ Each sentence was marked for each individual ▪ A total score was created for each individual	
Mean score	50% of each test	▪ A mean score of the two was created for each individual	

For the scoring of the EIT performances, an excel file was created with all of the sentences (30 in total) for each participant, and each sentence was marked individually. Then the total score of all the sentences was calculated using the relevant formula on Excel for each participant. Since the maximum scores that could be obtained from the EIT and the OPT were different (i.e. 120 and 100, respectively), it was necessary to convert either score into the other for the comparability of the scores as well creating a mean score to group the participants into different proficiency levels.

Initially, the EIT scores were converted into 100 and this was checked against the CEFR levels. However, it was observed that a holistic evaluation of the recordings for the participants' proficiency levels was better in line with the groupings based on a maximum score of 120 rather than 100. Therefore, the decision was taken to base the CEFR groupings on scoring out of 120. As such, the OPT scores were converted into 120, and then a mean score of the OPT and the EIT scores were calculated (50% of each). All these calculations were done using the relevant formulae on the excel file, and the mean scores ranged between 44 and 80. Here I should note that a Pearson correlational analysis was also carried out between the OPT and the EIT scores in order to understand how much these were related to each other. The results suggested that a high score on the OPT was associated with a high score on the EIT (r =.51, p < .001). The results also demonstrated that 26% of the variance in participants' OPT scores could be accounted for by their EIT scores.

Using Oxford Online Placement Test's (OOPT) scoring system, the participants were then grouped into the CEFR levels (please see OOPT website for further information) (Table 3.5). The OOPT is based on a scoring system of out of 120 and places participants into bands of 20. In other words, for each score above, each band of 20 equals to one CEFR level. For example, any score falling into the band of 21-40 belongs to B1 level at the CEFR. Accordingly, the participants of the current study were grouped into three CEFR levels; 15 at A2 level, 15 at B1 level and 12 at B2 level.

Table 3.5. The groupings of the participants across CEFR levels based on OOPT scoring system in Study 1

CEFR Levels	Mean score of OPT and EIT*	n (N=42)
C2	81>100	-
C1	61-80	-
B2	41-60	12
B1	21-40	15
A2	1-20	15

** The maximum score is 120.*

3.9.2.Fluency measures used in the study

As was mentioned in the previous chapter, there is currently not a consensus in fluency studies on which measures should be employed to examine fluency performance.

Although some studies, e.g. those comparing perceived fluency to utterance fluency, opted to use measures which corresponded raters' judgements (e.g. speech rate and mean length of run) (Prefontaine, Kormos, & Johnson, 2016), the findings from this line of studies have been controversial since such ratings are largely dependent on the instructions given to raters and equally important, on how fluency is defined in a given study (De Jong et al., 2013).

For a more systematic approach to fluency measurement, three key aspects of utterance fluency were put forth by Tavakoli and Skehan (2005); breakdown, repair and speed fluency, which were already explained in Section 2.4.1.2. In line with this framework, most fluency studies (e.g. De Jong et al., 2015; Huensch & Tracy-Ventura, 2016; Hunter, 2017; Tavakoli et al., 2016 to name a few) have examined utterance fluency using measures in these three sub-dimensions. Recent studies have pointed at a need for a more detailed approach to fluency measurement with a fourth category; i.e. composite measures (e.g. De Jong et al., 2015; Hunter, 2017; Tavakoli et al., 2019; Skehan, 2014). This body of research has suggested that some measures could be indicative of more than one aspect of fluency and therefore a distinction should be made between pure and composite measures. Examples of such composite measures include speech rate, which is a measure most commonly used as a speed measure in fluency studies but has been argued to be confounded with pausing behaviour, or mean length of run, which combines aspects of speed and breakdown fluency (Tavakoli, 2016) and has also been linked with perceived fluency (Kormos & Denes, 2004).

Accordingly, the choice of measures in the current study was made in a way that each fluency aspect (i.e. breakdown, repair and speed fluency) is represented with composite measures being included as a separate category. Given the exploratory nature of Study 1, a comprehensive choice of measures was needed for an in-depth analysis of fluency and a better understanding of the processes underlying L1 and L2 speech production. Therefore, unlike some previous studies examining L1 and L2 speech fluency (e.g. De Jong et al., 2015; Huensch &Tracy-Ventura, 2016; Riazantseva, 2001; Towell et al., 1996) which did not examine all aspects or employed only a few measures, a wide range of measures were adopted in Study 1 (17 in total), which are explained below.

3.9.2.1. Breakdown measures

It was explained earlier that pauses have been calculated rather differently by fluency researchers; e.g. in earlier studies 0.4 sec (Freed, 2000; Tavakoli & Skehan, 2005) or 0.1sec pause (Foster & Skehan, 1996) or in more recent studies 0.25 sec (e.g. De Jong et al., 2012b; De Jong & Bosker, 2013). In the current study, the cut-off point for pauses was chosen as 0.25 sec. since recent studies suggested that this is a noticeable length by native speakers of English and is a better indication of pausing behaviour (De Jong & Bosker, 2013). The pauses were examined in terms of their frequency (i.e. total number of occurrence), quality (i.e. filled or silent), location (i.e. mid-clause or end-clause) and length (i.e. duration of pauses). While pauses are most typically examined in terms of their frequency as it affects listeners' perceptions of fluency (Prefontaine & Kormos, 2016; Rossiter, 2009), pause length is also suggested to be reflective of breakdown fluency behaviour (Segalowitz et al., 2017). Although there is no consensus in studies whether length (Wang, 2014) or frequency (Lambert et al., 2017) is more indicative of pausing behaviour, pause length is examined in this study because how long a speaker pauses is also linked to processing difficulties encountered in speech production and thus examining pause length, in addition to pause frequency, would deepen our understanding of the processing mechanisms underlying L1 and L2 speech.

Another feature considered when examining pausing behaviour in the present study related to the location of pauses. There has been evidence from several recent studies that not only how frequently pauses occur but also where they are located within the speech are crucial to distinguish native and non-native speakers (Kahng, 2014; 2020 Skehan & Foster, 2007; Skehan et al., 2016). While L1 speakers tend to pause more often at clause-boundaries, L2 speakers do so within clauses. It is assumed that native speakers construct clauses one at a time without focusing on the structure (due to automatic processing); yet L2 speakers, depending upon their proficiency level, are engaged more consciously in message construction (due to their less automatized linguistic knowledge). In this regard, the location of pauses, rather than the frequency, might have an even greater impact on listeners' perceptions (De Jong, 2016; Suzuki & Kormos, 2019; Tavakoli, 2011). At this point, it is also worth noting that some previous studies (e.g. De Jong, 2016; De Jong et al., 2015; Huensch & Tracy-Ventura, 2016, 2017; Skehan & Foster, 2007) investigated pauses at AS-unit boundary rather than clauses. The present study, however, opted to examine pauses at clause boundaries drawing on the research evidence suggesting that the distinction between mid (i.e. within) and end (i.e. between) clause pauses provides a deeper insight into the differences between L1 and L2 fluency behaviour (e.g. Kahng, 2014; 2020; Tavakoli, 2011).

A final important feature pertains to the quality of pauses. Previous studies (Schmid & Fagersten, 2010) reported that the character of pauses - filled (i.e. non-lexical fillers such as erm, um, errr, etc.) and silent pauses- significantly distinguished fluent and disfluent speakers. Therefore, a consideration of both would be highly important for a better understanding of pause-related behaviour (De Jong & Perfetti, 2011).

Following from the above discussion, a total of twelve breakdown fluency measures were explored in the current study; i.e. total numbers and lengths of filled and silent pauses (both within and between clauses). A detailed description of these measures are provided in Table 3.6 below.

3.9.2.2. Repair measures

When it comes to measuring repair fluency, the approach adopted by SLA studies seems to be less clear. It may be due to the fact that several distinctions are made between types of repairs and these could impact upon listeners in different ways (Olynyk, d'Anglejan, & Sankoff, 1990). One such distinction is made between error repairs (i.e. pertaining to linguistic errors), and appropriateness repairs (i.e. occurring when speakers present a new or reformulated message (Kormos, 2006; Levelt, 1989). Another distinction is made between overt and covert repair (Kormos, 2006), which is more related to the way repairs are measured; while the overt repair can be identified though an analysis of the overt speech, i.e. the speech produced, the covert repair can be measured though the use of 'retrospective comments' on the speech (ibid.).

While exploring all repair types would be valuable, given that the focus of the current study is on the examination of utterance fluency (i.e. observable indices in the overt speech which are indicative of underlying processes in speech production), overt repair was examined here through the frequencies of repair instances; namely repetitions, replacements, reformulations and false starts. Such instances have been suggested to be examples of repair strategies used by L2 speakers (Hunter, 2017; Kormos, 1998; Skehan, 2003; 2009; 2015) and have been linked with cognitive and monitoring processes (Kormos, 1998; 2006). In other words, L2 speakers use these strategies when they engage in a repair process, possibly due to the demands in conceptualization (e.g. amount of information) or formulation (e.g. lexical retrieval or encoding). Since the present study is exploratory in terms of the measured used, all of these measures were examined here including the total number of these as a separate measure. As such, there were a total of five repair measures used: numbers of repetitions, replacements, reformulations, false starts and a total repair. Definitions of these are provided in Table 3.6 below.

3.9.2.3. Speed measures

Since the current study makes a distinction between pure and composite measures, the only speed fluency measure adopted is articulation rate. It is measured by calculating the total syllables produced and dividing it by total speaking time excluding the pause time, and is arguably the only measure that 'isolates the speed as a variable' (Hunter, 2017).

In the calculation of speed-related measures, one argument is about the basic unit of analysis; i.e. whether the calculations should be based on syllable counts or word counts. Once again, researchers adopt differing approaches to the issue, while some counting the words (Witton-Davies, 2014)) some others take syllable counts into account (e.g. Awwad, 2017; Huensch &Tracy-Ventura, 2016, 2017). I opted to adopt syllable counts as the basic unit of analysis for two reasons. Firstly, given the cross-linguistic nature of the study, words counts may be not be a reliable unit since the number of words are likely to change largely across the two languages under investigation. For example, one word in Turkish such as 'kitapliktan' would mean three words in English 'from the bookcase'. Secondly, most fluency research of this kind (e.g. De Jong et al., 2015; Huensch & Tracy-Ventura, 2016; 2017; Peltonen, 2018) have used speed measures based on syllable counts, and it would be ideal to be able compare the findings from such studies with the current ones in order to develop a better understanding of the speed fluency behaviour. Additionally, the calculation of fluency measures adopted here (e.g. articulation rate) requires syllable counts.

I should also note that the calculation of the articulation rate in this study was based on unpruned speech; i.e. all uttered syllables including the ones in repair instances were included in the analysis.

3.9.2.4. Composite measures

Three composite measures were used in the study; speech rate, phonation-time ratio and mean length of runs. Speech rate is a measure of the speed of speech delivery and is based on the number of syllables produced divided by total speaking time (including pausing time) in seconds (Kormos, 2006). It should be taken as a composite measure because mathematically speech rate is dependent on the number and duration of silent pauses and in this sense relates to the pausing behaviour. In other words, the more or longer a speaker pauses, the slower their speech rate will become (De Jong, 2016, p. 211). Thus, speech rate is not only indicative of the speed of speech delivery but about the extent to which a speaker pauses.

Following a similar line of reasoning, mean length of run is also taken as a composite measure. It indicates how lengthy the speech is between two pauses and is typically measured in terms of the average number of syllables of 'runs', which are utterances between two pauses (Kormos, 2006). The length of runs would give us an insight about the pausing behaviour in that longer runs would mean that there is less pausing happening within the clauses. In this sense, this measure is, too, sensitive to pausing behaviour.

The final composite measure is phonation-time ratio, which is the percentage of the time actually spent speaking (i.e. excluding pausing time). Some studies consider it as a breakdown measure (e.g. Hunter, 2017) claiming that it provides information about the amount of pausing. However, following from other studies (Tavakoli, 2016), phonation-time ratio was adopted as a composite measure in this study because while it considers the speed, it is dependent upon the amount of pausing. Table 3.6 below presents the definitions of all fluency measures used in Study 1.

Table 3.6. Fluency measures used in Study 1 (Kormos, 2006).

Breakdown Fluency Measures (pauses above 0.25 sec)	
Number of *mid-clause filled* pauses (MCFP) / minute	The total number of filled pauses within the clauses divided by the total speaking time
Length of *mid-clause filled* pauses (MCFP)/ minute	The total length of filled pauses within the clauses divided by the total number of mid-clause filled pauses
Number of *end-clause filled* pauses (ECFP)/ minute	The total number of filled pauses between the clauses divided by the total speaking time
Length of *end-clause filled* pauses (ECFP) /minute	The total length of filled pauses between clauses divided by the total number end-clause filled pauses
Number of *mid-clause silent* pauses (MCSP)/ minute	The total number of silent pauses within the clauses divided by the total speaking time
Length of *mid-clause silent* pauses (MCSP) /minute	The total length of silent pauses within the clauses divided by the total number of mid-clause silent pauses
Number of *end-clause silent* pauses (ECSP)/ minute	The total number of silent pauses between the clauses divided by the total speaking time
Length of *end-clause silent* pauses (ECSP) /minute	The total length of silent pauses between the clauses divided by the total number of end-clause silent pauses
Repair Fluency Measures	
Number of *repetitions*/ minute	The total number of words, phrases or clauses repeated with no modification
Number of *replacements*/ minute	Total number of phrases or clauses repeated with some modification
Number of *reformulations*/ minute	The total number of lexical items immediately substituted for another

Number of *false starts*/ minute	The total number of utterances that are abandoned before completion
Total repair	The total number of repair instances; i.e. repetitions, replacements, reformulations and false starts.
Speed Fluency measures	
Articulation rate	Number of syllables produced divided by total speaking time (excluding pause time) in seconds
Composite measures	
Speech rate	Number of syllables produced divided by total speaking time (including pause time) in seconds.
Phonation-time ratio	The percentage of time spent speaking as a percentage proportion of the time taken to produce the speech sample.
Mean length of runs	An average number of syllables produced in utterances between pauses of 0.25 seconds and above.

3.10. Data transcription and coding

In this section, I describe the ways in which the data collected through oral narrative tasks in both L1 and L2 were analysed. More specifically, these include how the data were transcribed, coded and annotated for the fluency measures adopted. In addition, I provide a detailed description of the data preparation process for statistical analysis with some specific examples from the actual data.

3.10.1. Simple transcriptions

As noted earlier, speech samples were elicited from the participants through audio-recordings of their performance on the narrative tasks. Initially, all of the recordings were transcribed using an online software, called Transcriber (2005). The majority of the participants produced speech shorter than 60 seconds, thus all of the speech samples were used for analysis. The Transcriber software included such features as slowing down, playback or rewinding and allowed for exporting the transcribed data into a word document. For each participant, a separate word document, including the simple transcriptions of L1 and L2 speech samples, was created. In the process of simple transcriptions, the following points were considered:

- All transcriptions started with the first utterance that the speaker produced. This means that any silences at the beginning of the recording were excluded from the analysis.
- No punctuation marks were used and all utterances, including filled pauses, were transcribed. In Turkish, occasional uses of fillers such as ‘ *ay’* ‘*sey*’ were also considered as filled pauses since these did not have any meaning in the context. In a similar vein, in the rare cases when the participants filled the silence with short laughs, these were treated as filled pauses as well, regardless of the language used.
- Only one word, i.e. erm, was used for all the filled pauses produced (e.g. umm, himm, ehm, uh, erm etc.).
- All contractions or reduced forms were considered as two words. For example, in English ‘*wanna*’ was taken as ‘*want to’* or ‘*don’t*’ as ‘*do not’*. Similarly in Turkish,

all omissions of vowels such as in ‘ *orda’*(there) were taken as the original word ‘ *orada’* or ‘*kaybolmak*’ (to get lost) as ‘*kayip olmak’*. The purpose here was to obtain a standardization in all the transcriptions.

- It was observed that in Turkish, a couple of the participants produced words in English, such as ‘*dog*’ instead of ‘*kopek*’ in Turkish. In these cases, the English words were translated into Turkish and the Turkish words were taken into consideration for analysis.

Once the simple transcriptions were completed, the word files were copied into a different folder and these new copies were used for further analysis. Keeping the simple transcriptions in this way would help the researcher at a later stage do the syllable counts and lexical complexity analysis for further analysis.

3.10.2. Coding for units of analysis: AS-units

The next step was to code the data into analysable speech units. Segmenting the data into analysable units is essential in SLA research so that the speech data could be analysed quantitatively (Foster, Tonkyn, & Wigglesworth, 2000). Researchers have used different units to date for this purpose. Some of the commonly used units included semantic units such as C-unit, defined as ‘utterances, for example, words, phrases, sentences, grammatical and ungrammatical, which provide referential or pragmatic meaning (Pica, Holliday, Lewis & Morgenthaler, 1989, p. 72) or syntactic units such as T-unit, which is essentially a main clause and clauses that are dependent on it (Hunt, 1970). Although these have been used in previous research of this kind (Iwashita, 2001; Kuiken & Vedder, 2008), such units have been suggested to have two main shortcomings: 1) they do not provide clear definitions for researchers or analysts; i.e. definitions vary or are ‘too simple to be used with real spoken data’ and 2) usually, few examples are provided in studies and these resemble little to the messy speech data researchers have to deal with, making it almost impossible to work on the real data in a reliable and systematic way (Foster et al., 2000, p. 357-363). Due to such shortcomings, it is also argued that researchers who analyse their data using these will both miss a lot of the data and find it hard to compare the results reliably.

In order to respond to these, Foster et al. (2000) developed AS-units. An AS-unit is defined as “a single speaker’s utterance consisting of *an independent clause or sub-clausal unit,* together with any *subordinate clause(s)* associated with either” (Foster et al., 2000, p. 365). This means that AS-units consist of one main clause and/or one or more connected clauses; if there is more than one clause, one unit should be identified as the main clause while the others as subordinate clauses. Two clauses, which are not connected with subordination or linking phrases (e.g. a relative clause, an adverb or a conjunction) should be classified as two separate AS-units.

Foster et al. (2000) suggest that using AS-units to segment the data into analysable units has several advantages over previously proposed units. For example, unlike a T-unit, an AS-unit considers any independent sub-clausal unit, and unlike a C-unit, it identifies the nature of any sub-clausal units. For example, an independent sub-clausal unit ‘can be elaborated to a full clause by means of recovery of ellipted elements from the context of the discourse or situation’ (ibid., p. 366) as in the example A: ‘where do you come from?, B: Madrid (which could be ‘ *I come from Madrid’*). Similarly, an elliptical utterance, such as ‘yes’, ‘oh’ ‘really, is also classified as an independent unit. Given that such expressions are common in speech, AS-units perform better in picturing the characteristics of spoken data. Thus, they have been widely accepted as the ideal way of dividing the raw and transcribed data into analysable units and have commonly been used in recent studies (e.g. Awwad & Tavakoli, 2019; De Jong et al., 2015; Hunter, 2017; Malicka & Levkina, 2012; Tavakoli & Foster, 2011 to name a few). Following from Foster et al. (2000) and such recent studies, the data in the current study were coded using the AS-units. The coding procedure required listening to the recordings for a few times to ensure that the AS units reflected the intonation patterns of speakers.

3.10.3. Classification of repair fluency

After the transcriptions were coded for their AS-units, repair instances were identified and marked with a set of conventional symbols (Appendix H). Repair fluency was categorised of repetitions, replacements, reformulations and false starts. The rationale for the choice of repair measures comes from previous studies (e.g. Hunter, 2017; Skehan, 2003, 2009, 2015) which suggest that these are the strategies usually used by L2 learners. The definition of each is provided in Table 3.6 and was also explained in Section 2.4.1.2.

Instances of each repair type were counted for each speech sample (in both languages), and then these numbers were averaged to 60 seconds for comparability since the total speech lengths in most samples were less than 60 seconds. In order to perform the calculations, the relevant formula in the excel file was used; e.g. total number of repetitions x 60 / total speech length in seconds. Inter-rater reliability for the classification and the placements of the repair type is discussed in Section 3.11 below.

3.10.4. Identification of pauses: using PRAAT

Filled pauses were already identified when the audio files were transcribed. However, silent pauses and lengths of all pauses (i.e. filled and silent) were identified later using PRAAT software (Boersma & Weenink, 2013). This temporal analysis, i.e. the timing of the speech and of the silences, complemented the data on the transcripts, and made it possible to calculate the speed and composite measures, which required the application of specific formulae using the pausing measures, i.e. breakdown measures.

PRAAT has been used by a number of researchers to assist in their analysis of L2 speech (Cucchiarini et al. 2000, 2002; De Jong & Perfetti, 2011; De Jong & Wempe, 2009). Growing in popularity, this software allows researchers to make an automatic analysis of certain measures (e.g. number or lengths of silent pauses). However, one of the limitations of the software is that not all fluency measures could be calculated automatically. For example, speech rate, which also considers filled pauses, cannot be measured by PRAAT. Furthermore, although PRAAT can provide a text grid with silences within the text, these still require a careful and detailed manual investigation since some background noises, e.g. page turns, could be marked as speech on the grid. Consequently, it requires a considerable investment of time for the researcher, especially considering the fact that 10% of the data need to be re-analysed using PRAAT for inter or intra-rater reliability (Section 3.11). Despite these limitations, however, the use of the software serves as an objective way to identify pauses and their lengths relatively accurately (Segalowitz, 2010), especially when dealing with larger amounts of data.

There were a total of 84 speech samples in the current study (i.e. 42 participants producing speech in both Turkish and English). As a first step in PRAAT analysis, each of these samples (i.e. audio files) was converted to .WAV sound format, which is congruent with PRAAT. Then, a text grid with segments was created for each file using the Annotate to Textgrid (silences) command. Tiers for filled pauses and silent pauses were also added to the grid to mark the beginnings and the endings of the pauses against the spectrogram (Figure 3.1). The analysis began at the first utterance produced by the speaker, be of this a filled pause or an actual word, and if a segment of silence in the beginning was cut out (because there was no utterance yet), the exact length of this segment was added right to the end of the file. For example, if the analysis began at 3.10 seconds, it ended at 43.10 seconds, resulting in the annotation of a total of 40 seconds, if the speech was less than 60 seconds.

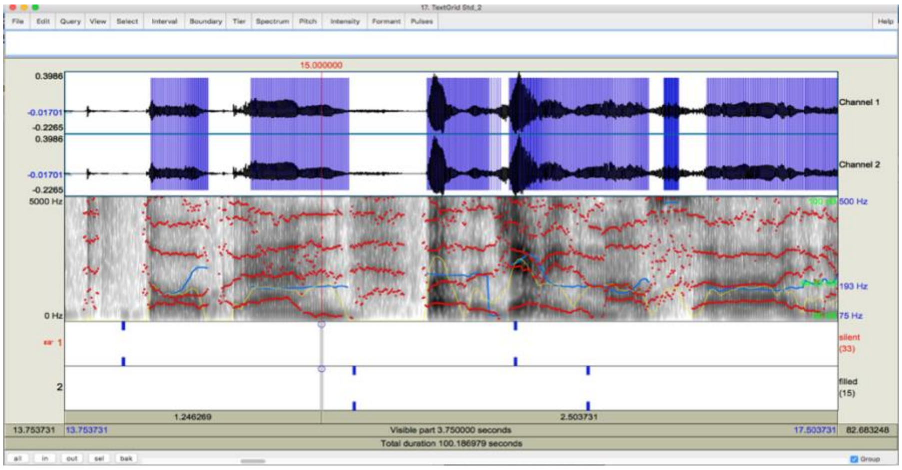


Figure 3.1. A PRAAT sound file and textgrid with accompanying tiers

Annotate to Textgrid (silences) command provides researchers with silent pauses (indicated as silences on the spectrogram) and the segments of utterances produced within the speech (indicated as sounding on the spectrogram). These are not always accurate though, as already explained. Therefore, one should zoom in multiple times and see additional information about the speech sample (e.g. intensity or pitch). Such additional information makes it easier for the researcher to identify the beginning and end of pauses, which enables very precise measurement.

Following from previous research (e.g. De Jong & Bosker, 2013; De Jong et al., 2015, Huensch & Tracy-Ventura, 2017; Kahng, 2014), silences of .25 of a second or longer were marked on the grid after multiple listening and checks for their location and the precise lengths. These, together with the total numbers of all pauses indicated on the tiers, were later noted down on each participant's word file to be used at a later step in the analysis for the identification of mid-clause and end-clause pauses (see an example of a coded sheet in Appendix H). Since PRAAT analysis provided with very precise measurements, be of exact length of speech in milliseconds or any exact words or syllables within 60 seconds (i.e. the cut off point for all recordings), at this phase, the coded sheets were re-visited again to complete or correct any of the missing or wrong information. In other words, the simple transcriptions along with the total speech times were double checked in this way.

3.10.5. Syllable counts

As was mentioned earlier, syllable counts were used in the current study as the basic unit of counting. To do the syllable counts, the simple transcriptions were revisited and the texts were cleared from any filled pauses or punctuation marks, if any. The syllables in Turkish speech samples were counted manually while a syllable counter website (syllablecount.com), following from Thai and Boers (2016), was used to do the counts in English speech samples. The counts in English samples were done by copying the text into the relevant column in the counter and using 'count syllables' command. In a different column, the counter provided the words that could not be identified or found in the dictionary, which also allowed the researcher do a more reliable check for any words that were not written correctly. The total numbers of syllables were noted down on the coded sheet of each participant for both languages.

In order to ensure that the counts were carried out accurately, 10% of the samples in English were double checked by manual counts. The agreement between the manual counts and those provided by the tool was above .98 (Cohen's Kappa). Also, syllables in 10% of Turkish samples were re-counted for intra-rater reliability, and a simple correlation of 1.00 between the first and second counts was achieved (also see Table 3.8). For all syllable counts, unpruned data, i.e. data including repair instances of repetitions, replacements, reformulations and false starts, were used.

3.10.6. Combining all scores for statistical analysis

The final step of the data analysis was to compile all analysis and measures onto one sheet to prepare the data for statistical analysis. During the entire process, most of the information needed for calculation of the fluency measures were already noted down on the coded sheets created initially for each individual; i.e. the number of AS-units and clauses, the number of repair instances, the precise lengths of all pauses and speech samples in both Turkish and English and number of syllables produced in each language. Further to these, numbers and lengths of mid-clause and end-clause pauses were calculated and again noted down on the same sheets (Table 3.7).

All of these calculations were transferred into the excel file, which had already contained the scores from proficiency tests of OPT and EIT from a previous stage of the analysis. On this file, the calculations of the remaining fluency measures were done using the relevant formula in the excel; e.g. for articulation rate, the number of syllables x 60 / total speech time (excluding pauses) in secs was used. Finally, all scores on the excel file were transferred into an SPSS file for further statistical analysis. Please see Table 3.7 below for a summary of all procedures followed in Study 1 regarding data transcription, data coding and data preparation for statistical analysis

Table 3.7. A summary of the data transcription and coding procedures in Study 1.

Stages of data preparation for statistical analysis	Steps followed
Fluency analysis	
Simple transcriptions	▪ **A word file (copy 1)** was created for each participant including L1 and L2 speech ▪ Speech samples were transcribed (Transciber (2005)) with filled pauses and no punctuation marks (*identification of filled pauses*) ▪ Simple transcription files (copy 1) copied into another folder **(copy 2)**
Coding for AS-units	▪ The data were segmented into AS-units on new copies **(copy 2/coded sheets)**
Classification of repair fluency	▪ Repetitions, replacements, reformulations and false starts were marked with a set of conventional symbols on the coded sheets (copy 2). (*identification of instances of repair fluency)*
Identification of pauses: using PRAAT	▪ 84 audio files (Turkish and English, 42 each) were converted into .WAV format ▪ A text grid was created for each audio file (Annotate to Textgrid (silences) command) ▪ Tiers were added against the spectrogram: filled and silent pauses ▪ The whole of the audio files was analysed: - Beginnings and endings of silent and filled pauses were marked . (*identification silent pauses and lengths of all pauses*) ▪ Coded sheets (copy 2) were re-visited to complete or correct any missing information ▪ Number and lengths of pauses were noted down on the coded sheet (copy 2).

	▪ Number and lengths of mid-clause and end-clause pauses (filled and silent) were calculated and noted down on the coded sheet (copy 2) (*identification of mid-clause and end-clause filled and silent pauses*)
Syllable counts	▪ Simple transcriptions were re-visited and copied into another folder (**copy 3)** ▪ New copies were cleared from any filled pauses ▪ Syllable counts were performed: - Syllablecount.com was used for English speech samples - Manual counts were used for Turkish speech samples ▪ Number of syllables were noted down on the coded sheets (copy 2).
Combining all scores for statistical analysis	
Transferring all the scores and measures onto the excel file and SPSS	▪ Coded sheets (copy 2) were re-visited and fluency measures were transferred into the excel file (Ex1). - Numbers and lengths of filled and silent pauses (mid-clause and end clause). - Number of repairs - Number of syllables - Total speech time ▪ Remaining fluency measures (e.g. articulation rate, speech rate) were calculated using the relevant formulae in Excel. *(identification/calculation of speech and composite fluency measures)* ▪ Fluency measures were approximated to 60 seconds, where applicable. ▪ Scores from proficiency tests were transferred into the excel file (Ex1). ▪ All scores were transferred into SPSS.

3.11. Inter-coder/rater reliability

In order to check the reliability of the coding and the analysis of the data, a sub-set of the whole data set (10%-20%) was re-analysed. The re-analysis took place at different phases of the data analysis. To start with, 10% of the OPT test sheets were re-marked by a second expert. There was a 100 agreement between the markers (Cohen's Kappa of 1.00). Then, the same marker scored 20% of the EIT performances after he was trained on the scoring system (Table 3.4). Pearson correlation coefficient of .98 was achieved between the sets of scores given by the two markers.

Next, a second bilingual, also an expert in the field, checked a sub-set of coded sheets (20%). Combining all of the data (i.e. the data from second coding), Cohen's Kappa's statistics were performed to determine the consistency between the coders on the identification and the placement of AS-unit, the clause boundaries, and all repairs combined (i.e. repetitions, replacements, reformulations and false starts). For each of these (e.g. clause boundary replacement), a reliability of above 0.95 was achieved (see Table 3.8 for each of these statistics). Further, in order to achieve the reliability of the manual syllable counts in Turkish, 10 % of the data was re-counted by the researcher and intra-rater reliability of 1.00 was achieved between the first counting and second counting. As for the syllable counts in English, which was carried out using an online tool, 10% of the data were double-checked manually by the researcher, and there was a 100% agreement between the researcher's and the tool's calculations (Cohen's Kappa of 1.00).

Finally, for the reliability of the identification of silent pauses and the lengths of all pauses using PRAAT software, 10% of the data were analysed a second time by the researcher to obtain *inter-coder reliability*. A simple correlation coefficient between the first coding and the second coding was calculated on the placement of the silent pauses and the identification of the lengths of the all pauses; for each of these a correlation of above .95 was achieved. Table 3.8 presents all the inter-rater/inter-coder reliability calculations across different phases of the data analysis.

Table 3.8. Inter-rater/coder reliability across different phases of the data analysis in Study 1

Segments of the data re-analysed/coded (in both languages where applicable)	**Second-rater/coder and reliability statistics**
The scoring of the OPT (10%)	▪ An expert in the field ▪ Inter-rater reliability: ▪ Cohen's Kappa of 1.00 (100%) combining the data
The scoring of the EIT performances (20%)	▪ A trained second rater ▪ Inter-rater reliability: ▪ Pearson correlation coefficient of .96 between the sets of scores
Coding of the transcriptions (20%) on *the identification* and *the placement* of: • AS-units • clause boundaries • filled pauses • repairs (i.e. repetitions, replacements, reformulations and false starts)	▪ A second bilingual expert in the field. ▪ Inter-coder reliability: - Combining the data, Cohen's Kappa for the identification of: o AS-units and clause boundaries (1.00) o filled pauses (1.00) o repairs (.96) - Combining all data, Cohen's Kappa for the placement of: o AS-units and clause boundaries (.97) o filled pauses (1.00) ▪ repairs (1.00)
Syllable counts (10%) - Manual counts for Turkish speech samples - Counts of the online tool for English speech samples	▪ The researcher (counting of the data a second time) ▪ Intra-rater reliability - Pearson correlation coefficient of 1.00 between first and second counting. ▪ The researcher (manual checks) ▪ Inter-rater reliability: ▪ Cohen's Kappa of 1.00 (100% agreement)
PRAAT analysis on (10%): - The identification and the placement of the silent pauses	▪ The researcher (analysis of the data a second time) ▪ Intra-rater reliability:

- The identification of the lengths of all pauses (i.e. silent and filled pauses)	- Pearson correlation coefficient between the first and second analysis for the *identification* of: ▪ silent pauses (1.00) ▪ lengths of silent pauses (0.97) ▪ lengths of filled pauses (0.96) ▪ Pearson correlation coefficient of 1.00 between the first and second analysis for the *placement* of silent pauses.

3.12. Summary of the chapter

In this chapter, I outlined the methodological approach taken and the data collection and analysis procedures followed in Study 1. Starting with the aims and the design of the study, I provided in-depth explanations for the tools employed together with their justifications . Then I described the ways the data were analysed; i.e. the marking procedures of the proficiency tests, how the data were transcribed and coded into analysable units, and provided clarifications about the calculations of measures, including the use of PRAAT software. Additionally, I discussed the fluency measures used in the study and provided justifications for employing them in the study. Finally, I mentioned how the reliability of the data were ensured and the data were prepared for statistical analysis.

CHAPTER 4. STATISTICAL ANALYSIS AND RESULTS: STUDY 1

4.1. Introduction

This chapter presents the details of statistical analyses run in order to answer the RQs of Study 1. The results are presented in relation to each RQ and under two main sections. The first section details how the data were cleaned from any errors and outliers and preliminary analysis carried out, including the descriptive statistics and the visuals for the assumptions for the inferential statistics. The second sections provides inferential statistics used to explore the relationship between variables of the study. These include Pearson-product moment correlations, partial correlations and multiple regressions.

As already mentioned in the previous chapter, Study 1 had a mixed between-within factorial design to investigate whether L1 and L2 fluency behaviours are related and if so, whether these relationships are mediated by an individual learner variable of L2 proficiency and an external variable of task structure. While L1 and L2 fluency performances (represented through a range of fluency measures) were dependent variables of the study, L2 proficiency, described with three levels but used as a continuous variable, and task structure with two levels were the independent variables. 17 fluency measures in total were used to operationalize L1 and L2 fluency behaviour and these were also used as continuous variables.

4.2. Data screening and cleaning for errors and/or outliers

It was mentioned earlier that the Excel file which had initially been created in the data preparation stage to combine all the scores, was transferred into an SPSS file for statistical analysis. Before carrying out any inferential analysis, the data were screened for any errors which might possibly have occurred in the data entry, or any possible outliers, values way above or below other scores (Pallant, 2013; Tabachnick & Fidell, 2013). The only error detected at this stage was the entry of total speech time for one participant; an initial examination of the data set revealed that one participant's score for articulation rate was well out of the range of other scores. In order to detect the reason for this, first the coded sheet for this particular participant's speech and then the text grid from PRAAT analysis were checked back again for any errors. It was found out that an error occurred when the length of speech sample was noted on the coded sheet, which resulted in an extreme value, and it was corrected in all data sets created.

In order to ensure that there were no more outliers, the values for both dependent and independent variables were checked by using the 'Explore' command in Descriptives in SPSS. An inspection of the boxplots revealed that there were a couple of outliers in English mean length of MCFP, ECFP and MCSP, and English phonation-time ratio. These outliers were checked back in the data set (i.e. in the coded sheets and excel files) to see whether their scores were genuine or not, but not just errors in the data entry. It was revealed that although some of the outliers had genuine scores, there were two further errors in the entry of two separate participants' scores; i.e. in the total pause lengths for mid-clause filled and silent pauses. These errors had subsequently resulted in different scores in pause-related measures, i.e. phonation-time ratio and articulation rate for these participants.

Once these errors were corrected, the next step was to compare 5% Trimmed Mean Values, for the remaining genuine scores, with the mean scores to see if the two values were similar or very different (Pallant, 2013). SPSS calculates trimmed values by removing the top and below 5 five percent of the cases, and therefore this comparison would reveal whether the outliers in the data set were having a lot of influence on the mean score for a given measure. Since the mean scores were not very different from each other, the decision was taken to retain these cases in the data file.

4.3. Preliminary analysis

Descriptive statistics were run for all fluency measures 1) to describe the basic characteristics of the sample, i.e. the variance in participants' performance and 2) to have first impressions about the group differences (i.e. between L1 and L2 fluency measures). Since all the measures were used as continuous variables, the 'Descriptives' command in SPPS was used, with 'summary' statistics such as mean, median and standard deviation. Further to the descriptive statistics, paired-samples t-tests were also carried out in order to see whether the group differences were statistically significant, and the effect sizes (Cohen, 1988) were calculated to assess the magnitude of the effect. It should be noted that the effect sizes were interpreted using the field-specific criteria in Plonsky and Oswald, (2014), who proposed benchmarks of d = .60 (small), d = 1.00 (medium), and d = 1.40 (large) for within-group comparisons.

Table 4.1 presents the results obtained from the descriptive statistics and paired-samples t-tests for the fluency measures in Turkish and English. As can be seen in the table, the figures for fluency measures in Turkish as L1 and English as L2 were very similar for *lengths of filled pauses* and *silent pauses*, both in terms of *mid-clause* and *end-clause pauses*, *number of end-clause filled pauses* and *number of false starts*. For a number of other measures, the figures seemed different in the two languages. These included several measures of breakdown, repair and speed measures. There was an increase in the mean scores from Turkish to English for breakdown measures of *number of mid-clause filled pauses* (M = 2.67, SD = 3.61 and M = 5.75, SD = 4.07), *number of mid-clause silent pauses* (M = 9.24, SD = 5.31 and M = 19.51, SD = 5.34) and for repair measures of *repetitions* (M = .31, SD = .67 and M = 1.40, SD = 2.09), *replacements* (M = .46, SD = .99 and M = . 65, SD = . 85), *reformulations* (M = .59, SD = 1.09 and M = .78, SD = 1.04) and *total repair* (M = 2.82, SD = 2.46 and M = 3.77 and SD = 3.07) The observed differences for the breakdown measures were statistically significant, with t = -5.79, df = 41, p < .001 for *number of mid-clause filled pause and* t = -10.93, df= 41, p < .001 for *number of mid-clause silent pauses*. Cohen's statistics indicated medium to large effect sizes for these measures; i.e. d= 0.80 and d =1.92, respectively (Plonsky & Oswald, 2014). As for the repair measures, however, while the differences were statistically significant for repetitions (t = -3.06, df = 41, p = < .004) and total repair (t = -2.07, df = 41, p < .044) with small to medium effect sizes (d = 0.70 and d = 0.34, respectively), the figures failed to show any statistically significance level (i.e. below p < . 05) for other repair measures.

As can be seen from the table, the descriptive statistics for speed and composite measures indicated a statistically significant decrease from Turkish to English; the figures in Turkish were higher than in English and the differences were statistically significant; with M = 378.52, SD = 59.51 and M = 198.38, SD= 45.46 (t = 18.87, df = 41, p < .001) for articulation rate, M = 224.16, SD = 53.48 and M = 79.42, SD = 31.72 (t = 17.23, df = 41, p < .001) for speech rate, M = 59.05, SD = 10.06 and M = 39.77, SD = 67.64 (t = 8.83, df = 41, p < .001) for phonation-time ratio, and M = 9.35, SD = 3.13 and M = 2.85, SD = .98 (t = 14.09, df = 41, p <. 001) for mean length of run. Cohen's statistics showed large effect sizes for all these measures except phonation-time ratio; the large effect sizes for articulation rate, speech rate and mean length of run were d = 3.40, d = 3.29 and d = 2.80, respectively while the effect size could be considered small for phonation-time ratio. Here, I should acknowledge the risk of Type 1 error; however, considering such low p values, having a Type 1 error is very unlikely in these cases. Therefore, it could be suggested that large effect sizes indicate that the differences between L1 and L2 fluency measures were important.

A general view of the data, i.e. means and standard deviations, suggest that the participants were overall more fluent in their L1 than in their L2. The lower mean scores for *number of mid-clause pauses* (both filled and silent) and repair measures of *repetitions* and *total repair,* and higher scores for *articulation rate*, *speech rate*, *phonation-time ratio* and *mean length of run* in Turkish indicate a higher degree of fluency. Also, the higher figure for *number of end-clause pauses,* both *silent and filled* in Turkish further suggest that speakers tend to pause more often at clause boundaries in their L1, which is a distinction often made between native and non-native speakers.

The descriptives were also run for proficiency scores; i.e. the EIT, OPT and mean scores of the two. As presented in Table 4.2 below, the EIT scores ranged between 9 and 105, with M = 36.38, SD = 26.51, the OPT scores between 37 and 88, with M = 64.93, SD =

14.25 and the mean proficiency scores between 25 and 93, with *M* = 50.90, *SD* = 17.97. These figures indicate that the participants performed better in the OPT, with a smaller variance of the scores, than they did in the EIT.

Table 4.1. Descriptive statistics for L1 and L2 fluency measures in Study 1

Fluency measures	Minimum	Maximum	Mean	SD	t	Sig. (2-tailed)	CI (95% for Mean)		Cohen's d
Breakdown measures									
Tur number of MCFP	.00	14.13	2.67	3.61	-5.79	.001	Lower Bound	1.54	0.80
							Upper Bound	3.80	
Eng number of MCFP	.00	12.65	5.75	4.07			Lower Bound	4.47	
							Upper Bound	7.02	
Tur length of MCFP	.00	2.29	.46	.56	-.61	.541	Lower Bound	.28	0.10
							Upper Bound	.63	
Eng length of MCFP	.00	1.50	.51	.32			Lower Bound	.41	
							Upper Bound	.61	
Tur number of ECFP	.00	12.38	3.68	3.71	.72	.471	Lower Bound	2.52	0.13
							Upper Bound	4.83	
Eng number of ECFP	.00	9.99	3.24	2.71			Lower Bound	2.39	
							Upper Bound	4.09	
Tur length of ECFP	.00	1.91	.53	.48	-.10	.914	Lower Bound	.38	0.02
							Upper Bound	.68	
Eng length of ECFP	.00	2.58	.54	.47			Lower Bound	.39	
							Upper Bound	.69	
Tur number of MCSP	.00	27.55	9.24	5.31		.001	Lower Bound	7.58	1.92
							Upper Bound	10.90	

Eng number of MCSP	10.00	32.60	19.51	5.34	-10.93		Lower Bound	17.85	
							Upper Bound	21.18	
Tur length of MCSP	.00	3.37	1.33	.71	.86	.394	Lower Bound	1.11	1.16
							Upper Bound	1.55	
Eng length of MCSP	.41	5.71	1.19	.94			Lower Bound	.89	
							Upper Bound	1.48	
Tur number of ECSP	8.11	24.75	15.34	4.71	6.35	.001	Lower Bound	13.87	1.42
							Upper Bound	16.81	
Eng number of ECSP	3.72	15.57	9.80	2.84			Lower Bound	8.91	
							Upper Bound	10.69	
Tur length of ECSP	.56	.3.56	1.76	.78	1.24	.222	Lower Bound	1.52	0.18
							Upper Bound	2.01	
Eng length of ECSP	.55	3.93	1.61	.85			Lower Bound	1.34	
							Upper Bound	1.88	
Repair measures									
Tur repetitions	.00	2.49	.31	.67	-3.06	.004	Lower Bound	.10	0.70
							Upper Bound	.21	
Eng repetitions	.00	8.56	1.40	2.09			Lower Bound	.74	
							Upper Bound	2.05	
Tur replacements	.00	5.04	.46	.99	-.96	.341	Lower Bound	.15	0.20
							Upper Bound	.77	
Eng replacements	.00	2.41	.65	.85			Lower Bound	.38	

							Upper Bound	.91	
Tur reformulations	.00	4.51	.59	1.09	-1.22	.228	Lower Bound	.59	0.17
							Upper Bound	.25	
Eng reformulations	.00	3.77	.78	1.04			Lower Bound	.46	
							Upper Bound	1.11	
Tur false starts	.00	8.19	1.44	1.86	1.23	.223	Lower Bound	.86	0.23
							Upper Bound	2.02	
Eng false starts	.00	5.19	1.08	1.17			Lower Bound	.71	
							Upper Bound	1.44	
Tur total repair	.00	9.03	2.82	2.46	-2.07	.044	Lower Bound	2.05	0.34
							Upper Bound	3.59	
Eng total repair	.00	14.98	3.77	3.07			Lower Bound	3.77	
							Upper Bound	2.81	
Speed measures									
Tur articulation rate	212.2 2	569.97	378.52	59.51	18.8 7	.001	Lower Bound	359.97	3.40
							Upper Bound	397.07	
Eng articulation rate	119.4 4	264.88	198.38	45.46			Lower Bound	188.86	
							Upper Bound	207.90	
Composite measures									
Tur speech rate	101.9 3	340.39	224.16	53.48	17.2 3	.001	Lower Bound	207.49	3.29
							Upper Bound	240.82	
Eng speech rate	31.43	170.30	79.42	31.72			Lower Bound	69.54	
							Upper Bound	89.31	

Tur phonation-time ratio	33.66	76.44	59.05	10.06	8.83	.001	Lower Bound	55.92	0.39
							Upper Bound	62.19	
Eng phonation-time ratio	12.41	66.58	39.77	67.64			Lower Bound	35.76	
							Upper Bound	43.79	
Tur mean length of run	4.50	20.29	9.35	3.13	14.09	.001	Lower Bound	8.37	2.80
							Upper Bound	10.32	
Eng mean length of run	1.50	5.48	2.85	.98			Lower Bound	2.55	
							Upper Bound	3.16	

Note: SD, Standard Deviation; L1, first language; L2 second/foreign language; MCFP, Mid clause filled pauses; MCSP, Mid-clause silent pauses; ECFP, End-clause filled pauses; ECSP, End-clause silent pauses

Table 4.2. Descriptive statistics for the OPT and the EIT scores in Study 1

	Minimum	**Maximum**	**Mean**	**Std. Deviation**	**CI (95% for Mean)**	
EIT scores	9.0	105.0	36.38	26.51	Lower Bound	28.12
					Upper Bound	44.64
OPT scores	37.0	88.0	64.93	14.25	Lower Bound	60.49
					Upper Bound	69.38
Mean scores	25.0	93.0	50.90	17.97	Lower Bound	45.30
					Upper Bound	56.50

4.4. Preliminary screening of data for statistical analysis

Prior to carrying out any inferential statistics, the assumptions for violations of normality, linearity and homogeneity were checked. A Shapiro-Wilk test was used in order to test whether the dependent variables (i.e. L1 and L2 fluency measures)and the independent variables (i.e. EIT, OPT and mean proficiency scores) could be assumed to be normally distributed. A Shapiro-Wilk test was carried out to test this, and the results showed that the data were normally distributed for most of the variables with $p < .05$. The Sig. value was less than .05 for several measures in Turkish and English. Since the violations of normality could be common in large data sets, Q-Q lots were inspected as a further step since these could portray a better picture of the sample in relation to the normality. The Q-Q plots for both fluency measures in Turkish and English and the proficiency scores indicated reasonably straight lines, suggesting that in general these variables were approximately normally distributed. Therefore, no further action was taken to transform the data (please see Appendix I).

Further, the assumptions of linearity and homoscedasticity were checked. Before carrying out any correlational analysis, it is recommended that scatterplots are obtained in order to see whether the variables in question (i.e. fluency measures in L1 and L2) are related in linear fashion (Pallant, 2013). When these were obtained, the figures suggested no violations of the assumptions of linearity and homoscedasticity, with scores roughly evenly spread. Therefore, running Pearson product-moment correlations was appropriate to explore relationships between fluency measures in both languages. The figures for Q-Q plots and scatterplots are provided in Appendix I and Appendix J for space limitations.

4.5. Pearson product-moment correlations

The first RQ of the study examined whether there was any relationship between L1 fluency and L2 fluency measures. Given that the current study employed a wide range of fluency measures in L1 and L2 (17 in total), initially, a factor analysis was seen necessary to explore the relationships. This technique would take the large set of variables and form a smaller number of components based on the inter-correlations between the variables (i.e. L1 and L2 measures) (Pallant, 2013). In this way, it would also allow the researcher to reduce the number of measures before using them in further statistical analysis (e.g. multiple regressions). However, as a first step, when the data were assessed for the suitability for the factor analysis (Pallant, 2013; Tabachnick & Fidell, 2013), it was found out that the sample size of the study was not adequate for this test (KMO way less than .06).

Therefore, without reducing the number of measures employed using a factor analysis at an initial phase, the relationships between L1 and L2 fluency behaviours were explored using simple correlational analysis. The Pearson product-moment correlations were run between fluency measures in both languages for all aspects of fluency separately; i.e. breakdown, repair, speed and composite. A simple bivariate (i.e. between two variables) correlation was deemed appropriate to answer the RQ, because 1) the variables of the study, i.e. L1 and L2 fluency measures, were interval level (continuous) variables, and 2) correlational analysis not only describes the relationships, if any, but it gives the direction and the strength of the linear relationships as well (Lowie & Seton, 2013; Pallant, 2013). The results of the Pearson correlations are provided in Table 4.3. Please note that the sizes of the value of Pearson correlations were interpreted using the field-specific criteria recommended in Plonsky and Oswald (2014), i.e. benchmarks of r = . 25 (small), r = . 40 (medium) and r = .60 (large)

As can be seen in Table 4.3, the results demonstrated moderate to strong positive correlations for a number of fluency measures. As regards to breakdown measures, moderate to strong positive correlations were observed between the two languages for *number of mid-clause filled pauses* (r = .60, p = .001), *number of end-clause filled pauses* (r = .30, p = .048), *number of mid-clause silent pauses* (r = .34, p = .024) and *length of end-clause silent pauses* (r = .51, p = .001). No significant correlations were observed for other breakdown measures of *number of end-clause silent pauses* and *lengths of mid-clause filled pauses end-clause filled pauses* and *mid-clause silent pauses*.

As for repair measures, the correlations indicated positive moderate correlations for *number of reformulations* (r = .54, p = .001) and *total repair* (r = .44, p < .003) while no correlations were observed *for number repetitions*, *number of replacements* and *false starts*. Regarding other aspects of fluency, i.e. speed and composite measures, no significant correlations were found for any of these measures.

Table 4.3. Correlations between L1 and L2 fluency measures for all groups in Study 1

Fluency Measures	***r***	***p***
	Breakdown fluency	
Number of MCFP	.606**	.001
Length of MCFP	.219	.163
Number of ECFP	.306*	.048
Length of ECFP	.020	.899
Number of MCSP	.348*	.024
Length of MCSP	.163	.302
Number of ECSP	-.056	.727
Length of ECSP	.517**	.001
	Repair fluency	
Number of repetitions	-.145	.358
Number of replacements	.113	.475
Number of reformulations	.541**	.001
Number of false starts	.285	.067
Total repair	.447**	.003
	Speed fluency	
Articulation rate	.179	.257
	Composite measures	
Speech Rate	.266	.088
Phonation-time ratio	.260	.097
Mean length of runs	.301	.053

**Correlation is significant at the 0.05 level (2-tailed)*

***Correlation is significant at the 0.01 level (2-tailed)*

Interestingly, some L1 measures also correlated with some L2 measures other than their equivalents in the correlation matrix; for example, there was a significant correlation between *Turkish length of end-clause silent pauses* and *English length of mid-clause silent pauses* (r = .54, p =.001), and between *Turkish mean length of mid-clause filled pauses* and *English number of mid-clause filled pauses* (r = .41, p = .007).

Further to the Pearson correlations, coefficients of determination (i.e. square of r values) were calculated in order to see how much variance was shared by L1 and L2 fluency measures. Following the procedures described in Pallant (2013), the *r* values were multiplied by themselves first, and the resulting numbers were further multiplied by 100 to convert them into percentage values (Pallant, 2013). For example, L1 and L2 measures for number of mid-clause filled pauses correlate at r= .606 and share (.606 x .606) x 100 = 36.72 percent of their variance. This suggests that number of mid-clause filled pauses in L1 helps to explain nearly 37 percent of the variance in number of mid-clause filled pauses in L2, which is quite a reasonable amount of variance explained. Table 4.4 below presents R squared values for the significant correlations found in the Pearson-correlational analysis previously reported.

Table 4.4. Coefficients of determination for the significant results in the correlational analysis in Study 1.

Fluency measures	**R squared**
Number of MCFP	30.72
Number of ECFP	09.36
Number of MCSP	12.11
Length of ECSP	26.72
Number of reformulations	29.26
Total repair	19.98

As can be seen in the table, the amount of variance shared by L1 and L2 breakdown measures was 30.72% for *number of mid-clause filled pauses*, 9.36% for *number of end-clause filled pauses*, 12.11% for *number of mid-clause silent pauses* and 26.72% for *length of end-clause silent pauses*. Regarding the significant correlations found in repair fluency measures, namely reformulations and total repair , 29.26% and 19.98% of the variance were shared by L1 and L2 measures for these measures, respectively. Following from Plonsky and Ghanbar (2018), who proposed the benchmarks of .20 (or below) as small and .50 (or above) as large, these figures could be considered respectable amounts of variance in English fluency behaviour explained by Turkish fluency behaviour.

It was also explored whether the significant correlations observed in the Pearson correlations were similar across different proficiency groups. To analyse the correlations for each group separately, it was necessary first to split the data file and then to repeat the analysis. Therefore, after the data file was split, using Pearson product-moment correlations once again, the relationships between Turkish and English fluency measures were investigated across the groups. Table 4.5 below presents the results from Pearson correlations across different proficiency levels (i.e. A2, B1 and B2 levels).

Table 4.5. Correlations between L1 and L2 fluency measures across different proficiency levels in Study 1.

Fluency measures	A2 (n=15)		B1 (n=15)		B2 (n=12)	
	r	***p***	***r***	***p***	***r***	***p***
Breakdown fluency						
Number of MCFP	.292	.290	.552*	.033	.844**	.001
Number of ECFP	.389	.152	.347	.206	-.076	.813
Number of MCSP	.629*	.012	.171	.541	.246	.440
Length of ECSP	.536*	.040	.410	.129	.799**	.002
Repair fluency						
Number of reformulations	-.144	.609	.563*	.029	.547	.066
Total repair	-.045	.875	.595*	.019	.336	.286

**Correlation is significant at the 0.05 level (2-tailed)*

***Correlation is significant at the 0.01 level (2-tailed*

As can be seen in the table, there were significant positive correlations between L1 and L2 fluency measures for all proficiency groups of A2, B1 and B2 levels. There were moderate to strong positive correlations for *number of mid-clause filled pauses* in B1 and B2 groups (r = . 55, p < .033 and r = .84, p < .001, respectively), while the results did not indicate any correlations for *number of end-clause filled pauses* when the proficiency groups were analysed separately. L1 and L2 measures for *number of mid-clause silent pauses* correlated only in the lowest proficiency group, i.e. A2 (r = 62, p < .012). As for the length of end-clause silent pauses, moderate to strong correlations were observed for the lowest (i.e. A2) and highest proficiency (i.e. B2) groups only (r = 53, p < . 040 and r = 79, p < .002, respectively), but not for B1. Interestingly, significant correlations found for repair measures were maintained only for B1 level (r = 56, p < .029 for *number of reformulations* and r = 59, p < .019 for *total repair*). However, given that the group numbers for proficiency levels were relatively small, these results should be approached with caution.

4.6. Partial correlations

RQs 1(a) and 1(b) examined the extent to which the relationship between L1 and L2 fluency measures were moderated by the individual learner variable of L2 proficiency (i.e. mean scores) and an external factor of task structure, respectively. In order to address these questions, partial correlations were run separately with each of the two independent variables as the controlling factor. Partial correlations were used in the current study because it allows the researcher to statistically control for the influence of confounding variables (Pallant, 2013). Therefore, by removing the influence of L2 proficiency and task structure in the current study, a clearer and more accurate picture of the relationship between the L1 and L2 fluency measures could be created. The results from this statistical analysis would also make it possible for us to understand to what extent the relationships were influenced by these variables. If the influence of any of these is big, then we would expect to obtain a smaller correlation between the two measures than the zero-order correlations (i.e. significant correlations from Pearson correlations). In such a case, a smaller correlation would suggest that a weaker relationship exists between the two languages for a given measure.

4.6.1. Partial correlations controlling for L2 proficiency

In the previous section, simple bivariate correlations were carried out across the proficiency levels (i.e. B1, B2 and C1 levels) in order to explore whether the correlations observed between L1 and L2 fluency measures were similar in different proficiency groups. In this section, in order to answer RQ 1(b) (i.e. to investigate whether L2 proficiency mediates L1-L2 fluency relationships), the significant correlation coefficients between the measures in the two languages were further examined by running partial correlations, with L2 proficiency (i.e. the mean proficiency score of OPT and EIT scores) as the controlling factor. The results from the partial correlations controlling for L2 proficiency are presented in Table 4.6 below, where the zero order correlations indicate the Pearson correlations obtained in previous section and the r values indicate the results of the same analyses but this time taking out the effects of L2 proficiency.

Table 4.6. Partial correlations for the significant results, controlling for language proficiency in Study 1

Fluency measures	***r***	***p***	**Zero order (Pearson) correlation**
Number of MCFP	.606	.001	. 606, $p < .001$
Number of ECFP	.313	.046	.306, $p < .048$
Number of MCSP	.329	.036	.348, $p < .024$
Length of ECSP	.547	.001	.517, $p < .001$
Number of reformulations	.518	.001	.541, $p < .001$
Total repair	.446	.002	.447, $p < .003$

As can be understood from the table, the results indicated that moderate to strong correlations found between Turkish and English measures were maintained for all the significant results. An inspection of the zero order correlations suggested that controlling for language proficiency had little impact on the strength of relationship for *number of mid-clause filled pauses* (r = . 60, $p < .001$*), number of end-clause filled pauses* (r = 31, $p < .046$), *number of mid-clause silent pauses* ($r = .32$, $p < .036$), and *total repair* ($r = .44$, $p < .002$); with zero order correlations of r = 60, $p < .001$; $r = .30$, $p < .048$; $r = .34$, $p <$ .024 and $r = .44$, $p < .003$, respectively). Only for *length of end-clause silent pauses* and *number of reformulations*, a small change was observed in the strength of the correlation; a small decrease for the former (from .54 to 51) and a small increase for the latter (from 51 to 54), suggesting that L2 proficiency had a very little influence on the strength of the L1-L2 relationships for these measures. The results, when taken together, suggest that the observed relationships between L1 and L2 fluency measures are not due to the influence of language proficiency, but persist regardless of it at these levels.

4.6.2.Partial correlations controlling for task structure

In order to explore whether the relationships between L1 and L2 fluency behaviours are mediated by task structure, a second set of partial correlations were run for the same measures (i.e. the significant correlation coefficients from the Pearson correlational analysis, this time statistically controlling for the effect of task structure. The results demonstrated that this variable had very little impact on the strength of the relationships between L1 and L2 fluency measures. When partial correlations were compared with zero order correlations, it can be seen that the significant results were maintained for all fluency measures; for breakdown measures of *number of mid-clause filled pauses* (r = .60, p < .001*),number of end-clause filled pauses (r* = .33, p < .033*), number of mid-clause silent pauses* (r = .35, p < .025) and *length of end-clause silent pauses* (r = .51, p < .001) (with zero order correlations r = .60, p < .001, r = .30, p < .048, r = .34, p < .024 and r = .51, p < .001, respectively), and repair measures of *number of reformulations* (r = .54, p < .001) and *total repair* (r = .43, p < .004) (with zero order correlations, r = .54, p < .001 for the former, and r = .44, p < .003 for the latter). These results suggest that task structure did not have any influence on the strength of the fluency relationships between Turkish and English; i.e. the observed relationships between L1 and L2 were not due to the effect of this external factor. Table 4.7 below illustrates the results obtained from partial correlations controlling for this variable.

Table 4.7. Partial correlations for the significant results, controlling for task structure in Study 1

Fluency measures	***r***	***p***	**Zero order (Pearson) correlation**
Number of MCFP	.607	.001	.606 , p < .001
Number of ECFP	.335	.033	.306, p < .048
Number of MCSP	.350	.025	.348, p < .024
Length of ECSP	.510	.001	.517, p < .001
Number of reformulations	.542	.001	.541, p < .001
Total repair	.435	.004	.447, p < .003

4.7. Multiple regressions

RQ2 examined the extent to which measures of L2 fluency can be predicted from measures of L1 fluency. In order to address this question, standard multiple regressions were run for all fluency measures with L2 fluency measure as the dependent variable and L1 fluency measure and L2 proficiency measured by OPT and EIT scores as in dependent variables. Although this RQ is primarily concerned with the predictive power of L1 measures over L2 measures, L2 proficiency scores (i.e. OPT and EIT scores) were also added as other independent variables to the regression analysis to see to what extent they can predict L2 fluency measures.

Multiple regression was used in the current study to address the RQ2, because this analysis allows the researcher to explore a relationship between one continuous variable (e.g. L2 fluency measures in this case) and a number of other independent variables (e.g. L1 fluency measures and L2 proficiency scores). Though it can basically be considered another type of correlation, it was deemed necessary to conduct both analysis for different purposes. That is to say, while a correlational analysis was necessary to explore the degree of the relationship between two or more variables (i.e. L1 and L2 fluency measures) and to have a more concise and clear summary about it, regression analysis would show how one variable affects another variable. Regression analysis would enable a more detailed exploration of the relationship between a set of variables (Pallant, 2013, p. 134) and of how well these variables are able to *predict* or *estimate* a particular outcome by optimizing the data. Further, when a set of independent variables (i.e. L1 fluency measures and proficiency scores) are entered into a regression analysis, information can be obtained about the model as a whole as well as about the contribution of each of variables to the model; thereby the analysis indicates which particular variable is the best predictor of the model.

One issue that should be mentioned when using the regression analysis concerns the sample size of the study. It is suggested that when using multiple regressions, the results from a small data set may not be generalized or repeated with other samples (Tabachnick & Fidell, 2013) . Therefore, in order to estimate the power of the analysis for the sample size of the current study, GPower (Faul, Erdfelder, Buchner, & Lang, 2009) was used to run a post-hoc power analysis. The analysis was carried out for a linear regression fixed model to calculate the power of each individually significant regression model with an alpha level of 0.05, an effect size of 0.15 and a sample size of 42. The results showed a power of above .80 for all measures; .99 for *number of mid-clause filled pauses*, *length of end-clause silent pauses*, *speech rate*, *phonation-time ratio* and *mean length of run*s; .84 for *number of mid-clause silent pauses*; .98 for *number of reformulations* and .95 for *total repair*. The results of the power analysis suggested that a reliable level of confidence could be maintained in the findings from linear multiple regression analysis.

Table 4.8. Multiple regressions models predicting L2 fluency from L1 fluency measures, the OPT and the EIT scores in Study 1

Measures	**Test Statistics**	**B**	**SE**	**β**	***F***	***p***	**R Squared**	**Adjusted R^2**	**Effect size (Cohen's f^2)**
					Breakdown fluency				
Number of MCFP	Intercept	.690	2.42		F (3, 38) = 9.10	.001	.41	.37	.69
	L1 measure	.606	.146	.538		.001			
	OPT score	.072	.043	.253		.100			
	EIT score	-.035	.023	-.225		.138			
Length of MCFP	Intercept	.677	.243		F (3, 38) = 1.45	.243	.10	.03	.11
	L1 measure	.145	.091	.249		.121			
	OPT score	-.005	.004	-.227		.223			
	EIT score	.003	.002	.251		.171			
Number of ECFP	Intercept	5.75	1.92		F (3, 38) = 2.63	.064	.17	.10	.20
	L1 measure	.262	.115	.359		.028			
	OPT score	-.053	.035	-.277		.136			
	EIT score	-.001	.018	-.014		.938			
Length of ECFP	Intercept	1.031	.362		F (3, 38) = .722	.545	.05	-.02	..05
	L1 measure	.049	.158	.150		.756			
	OPT score	-.008	.006	-.240		.207			
	EIT score	.000	.003	.013		.945			
Number of MCSP	Intercept	9.06	4.33		F (3, 38) = 4.11	.013	.24	.18	.31
	L1 measure	.391	.149	.389		.012			
	OPT score	.143	.063	.382		.030			
	EIT score	-.068	.033	-.336		.049			

Length of MCSP	Intercept	1.94	.742		F (3, 38) = 1.49	.232	.10	.03	.11
	L1 measure	.223	.203	.168		.280			
	OPT score	-.014	.012	-.215		.238			
	EIT score	-.004	.006	-.101		.577			
Number of ECSP	Intercept	9.52	2.33		F (3, 38) = .758	.525	.05	-.01	.05
	L1 measure	-.021	.100	-.035		.836			
	OPT score	-.005	.039	-.027		.889			
	EIT score	.026	.020	.245		.202			
Length of ECSP	Intercept	2.42	.558		F (3, 38) = 10.62	.001	.45	.41	.81
	L1 measure	.530	.131	.486		.001			
	OPT score	-.029	.008	-.483		.001			
	EIT score	.004	.005	.110		.441			
				Repair fluency					
Number of reformulations	Intercept	-1.01	.653		F (3, 38) = 7.86	.001	.38	.33	.61
	L1 measure	.414	.129	.435		.003			
	OPT score	.025	.011	.349		.033			
	EIT score	-.003	.006	-.069		.647			
Number of repetitions	Intercept	-.107	1.57		F (3, 38) = 1.77	.169	.12	.05	.13
	L1 measure	-.353	.480	-.113		.467			
	OPT score	.041	.026	.277		.130			
	EIT score	-.028	.014	-.359		.051			

Number of replacements	Intercept	-.425	.629		F (3, 38) = 1.43	.248	.10	.03	.11
	L1 measure	.038	.138	.044		.785			
	OPT score	.015	.011	.243		.185			
	EIT score	.003	.006	.099		.595			
Number of false starts	Intercept	1.22	.859		F (3, 38) = 1.82	.159	.12	.05	.13
	L1 measure	.198	.099	.317		.052			
	OPT score	-.002	.015	-.018		.920			
	EIT score	-.009	.008	-.203		.258			
Total repair	Intercept	-.411	1.99		F (3,38) = 5.88	.002	.31	.26	.44
	L1 measure	.522	.182	.419		.007			
	OPT score	.066	.036	.308		.070			
	EIT score	-.044	.018	-.380		.020			
				Speed fluency					
Articulation rate	Intercept	176.22	37.72		F (3, 38) = 2.24	.098	.15	.08	.17
	L1 measure	.122	.079	.131		.130			
	OPT score	-.624	.374	.004		.104			
	EIT score	.456	.205	.275		.032			
				Composite measures					
Speech rate	Intercept	-.859	22.07		F (3, 38) = 11.70	.001	.48	.43	.92
	L1 measure	.153	.070	.258		.036			
	OPT score	.340	.309	.153		.277			
	EIT score	.655	.164	.547		.001			

Phonation-time ratio	Intercept	4.69	9.90		F (3, 38) = 13.66	.001	.51	.48	1.04
	L1measure	.130	.149	.102		.390			
	OPT score	.297	.122	.328		.020			
	EIT score	.224	.064	.461		.001			
Mean length of runs	Intercept	1.13	.659		F (3, 38) = 9.70	.001	.43	.38	.75
	L1 measure	.069	.039	.219		.084			
	OPT score	.005	.010	.074		.608			
	EIT score	.020	.005	.550		.001			

As indicated in Table 4.8, the results imply that for a number of measures, models predicting L2 fluency from L1 fluency and L2 proficiency scores reached a statistically significant level. As for breakdown fluency, the significant models were *number of mid-clause filled pauses*, F (3,38) = 9.10, $p < .001$, *number of mid-clause silent pauses*, F (3,38) = 4.11, $p < .013$, and *length of end-clause silent pauses* F (3,38) = 10.62, $p < .001$. The amounts of variance explained by these measures were 37% for number of mid-clause filled pauses, 18% for number of mid-clause silent pauses and 41% for length of end-clause silent pauses (adjusted R^2 = .41, .24 and .45, respectively). The model for *number of end-clause filled pauses* F (3,38) = 2.63, $p < .06$ just missed reaching a significant level, although 10% of the variance in the number of end-clause filled pauses was explained by this model (adjusted R^2 = .17). For number of mid-clause filled pauses, Turkish fluency measure significantly contributed to the model ($p < .001$) while language proficiency scores did not. For number of mid-clause silent pauses, all measures (i.e. L1 measure, the OPT and the EIT scores) significantly contributed to the model ($p < .012$, $p < .03$ and $p < .049$, respectively); however, for length of end-clause silent pauses only Turkish fluency measure and the OPT scores made significant contributions ($p < .001$ and $p < .001$, respectively) to the model while the EIT scores did not.

As for the repair measures, the models reaching a significant level were *number of reformulations* F (3,38) = 7.86, $p < .001$ and *total repair*, F (3,38) = 5.88, $p < .002$, with Turkish fluency measure contributing significantly to both models ($p < .003$ and $p < .007$, respectively). While the OPT scores made a significant contribution to the model for number of reformulations ($p < .033$) but not for total repair, the EIT scored contributed significantly to the model for total repair ($p < .020$) but not for number of reformulations. The amounts of variance explained by these measures were 33% for number of reformulations and 26% for total repair (adjusted R^2 = .38 and .31, respectively).

The only model for speed fluency, i.e. articulation rate, did not reach a statistically significant level. All the models for composite measures, i.e. speech rate, phonation-time ratio and mean length of runs, reached a statistically significant level, F (3,38) =

11.70, $p < .001$; F (3,38) = 13.66, $p < .001$, and F (3,38) = 9.70, $p < .001$, respectively. L1 measure contributed significantly only to the model for speech rate ($p < .036$). Interestingly, the contribution made by the OPT scores to the models was non-significant, while the EIT scores contributed significantly to all three models; i.e. for speech rate ($p < .001$), phonation-time ratio ($p < .001$) and mean length of runs ($p < .001$). The amounts of variance explained by these measures were 43% for speech rate, 48% for phonation-time ratio and 38% for mean length of runs (with adjusted $R^2 = .48$, .51 and .43, respectively). The results suggest that the composite measures of fluency can be predicted from language proficiency assessed by the EIT scores. To interpret the strength of adjusted R^2 values (for the whole models), Plonsky and Ghanbar's (2018) recent proposal was followed; i.e. values of up to .20 were considered small and those above .50 were regarded as large. Accordingly, all the R^2 values obtained for the significant models ranged between a small level of .18 and a medium level of .48, which indicates that the models were able to explain up to 48% of the variance in learners' performance.

4.8. Summary of key findings

In this chapter, I have presented the results from the statistical analysis carried out to address the RQs of Study 1. Before describing the statistical analysis in relation to each RQ, the preliminary analysis, i.e. descriptive statistics, and the procedures to check the assumptions of the statistical analysis were explained in depth. Then, the findings from the inferential analysis run to answer each RQ was presented with relevant tables. The two RQs that guided Study 1 and the summary of the findings in relation to these are provided in the table below (Table 4.9).

Table 4.9. A summary of the research questions and the key findings in Study 1.

	Research Questions	Overall Findings
RQ1)	To what extent are L1 Turkish and L2 English fluency behaviours related among lower-proficiency L2 learners in an EFL context?	To a certain extent. Significant correlations were found between L1 and L2 fluency measures for aspects of breakdown and repair. Moderate to strong positive correlations were observed between L1 and L2 fluency measures: - number of mid-clause filled pauses (*r* = .60) - number of end-clause filled pauses (*r* =30) - number of mid-clause silent pauses (*r* = 34) - length of end-clause silent pauses (*r* = .51) - number of reformulations (*r* = .54) - total repair (*r* =. 44)
RQ1a)	Does level of L2 proficiency, measured through OPT and EIT, mediate the relationship between L1 and L2 fluency behaviours among low-proficiency learners in an EFL context?	Overall, no impact of L2 proficiency was observed on the strength of the relationships between L1 and L2 fluency measures.

RQ1b)	Does task structure, i.e. tight and loose structure, mediate the relationship between L1 and L2 fluency behaviours among low-proficiency learners in an EFL context?	Overall, no impact of task structure was observed on the strength of the relationships between L1 and L2 fluency measures.
RQ2)	To what extent can L2 fluency behaviour be predicted from both L1 fluency behaviour and language proficiency scores among lower-proficiency learners in an EFL context?	To a certain extent. L1 measures predicted some of the measures in different fluency aspects. Turkish measures contributed significantly to the models for these fluency measures: - number of mid-clause filled pauses - number of mid-clause silent pauses - length of end-clause silent pauses - number of reformulations - total repair L2 proficiency scores made significant contributions to the models for these fluency measures: The OPT scores - number of mid-clause silent pauses - length of end-clause silent pauses - number of reformulations The EIT scores - number of mid-clause silent pauses - total repair - speech rate - phonation-time ratio - mean length of runs

CHAPTER 5. DISCUSSION: STUDY 1

5.1. Introduction

In this chapter, I attempt to interpret and discuss the findings of Study 1 in the light of the relevant literature and the findings reported in previous studies. The chapter is organized into two main sections; I start with an overview of the key findings from Study 1 and then discuss these under four further sub-sections which are 1) overall differences between L1 and L2 fluency measures, 2) correlations found between the two in general and across different proficiency levels, 3) partial correlations for the significant findings from simple correlations, controlling for L2 proficiency and task structure, and 4) multiple regressions predicting L2 fluency for all fluency measures from L1 fluency and L2 proficiency scores. The chapter then concludes with a summary of the key points discussed throughout.

5.2. Overview of findings from study 1

The prime aim of Study 1 was to investigate whether L1 and L2 fluency behaviours were related for learners of low-proficiency levels (i.e. A2, B1 and B2) in an EFL context. The study was motivated by recent studies which suggested that L2 speech fluency might, to a certain extent, mirror speakers' personal styles in their L1 (e.g. De Jong et al., 2015; Huensch & Tracy-Ventura, 2016; Peltonen, 2018). Related to this, there has been discussion around whether L2 speech fluency is a characteristic specific to an individual or a certain language being spoken; in other words, it has been questioned whether speech fluency is a trait or a language specific state (e.g. Derwing et al., 2009; Segalowitz, 2010) as some features of speech, such as personal styles, have been reported to be influential on listeners' perceptions, and ultimately their judgements about L2 fluency. Clearly, if we are to obtain a more reliable assessment of L2 learners' fluency, sources of such variability in L2 speech, which are not related to L2, should be identified and their effects should be removed from speech. Only then, it would be possible to look into L2-specific aspects only in the performance (De Jong et al., 2015; Segalowitz, 2010). Study 1, therefore, aimed to explore one such source, i.e. the influence of L1 speaking styles on L2 fluency performance by examining the relationship between L1 and L2 fluency behaviours.

By collecting the speech data in L1 Turkish and L2 English from a group of 42 undergraduate students in Turkey, performances in both languages were analysed for a wide range of fluency measures; i.e. a total of 17 fluency measures, including eight breakdown measures (numbers of mid-clause and end-clause filled pauses, numbers of mid-clause and end-clause silent pauses, lengths of mid-clause and end-clause filled pauses, and lengths of mid-clause and end-clause silent pauses), five repair measures (numbers of repetitions, replacements, reformulations, false starts and total repair), one speed measure (articulation rate) and three composite measures (speech rate, phonation-time ratio and mean length of runs) which combine the breakdown and speed aspects of fluency. The participants' proficiency levels were measured through two standardized tests; i.e. the OPT and the EIT. Related to L1-L2 fluency relationships, a further aim of Study 1 was to explore whether variations in L2 proficiency and task structure mediated factors in determining the strength of the relationships between L1 and L2 fluency behaviours. To achieve these aims, one research question with two sub-questions was formed and answered in the previous chapter. The results showed that L2 breakdown and repair fluency were to some extent related to L1 personal speaking styles. It was also found that overall L1-L2 relationships persisted regardless of the variability in L2 proficiency and task structure.

The secondary aim of the study was to explore the extent to which L2 fluency measures could be predicted from L1 fluency measures. It was also interested in examining the predictive power of L2 proficiency, i.e. the OPT and EIT scores, over L2 fluency. The statistical analysis carried out to answer this question were provided in the previous chapter. The results demonstrated that all aspects of L2 fluency could be predicted from L1 fluency, at least to some extent. While L1 measures made significant contributions to the models mostly for breakdown fluency and partly for repair fluency, the OPT and EIT scores predicted different aspects of the performance. The findings suggested that performance in OPT could partly predict mid-clause pausing while performance in EIT predict speed-related fluency, i.e. all composite measures which are closely related with speed aspect. In the following sections, I discuss these findings in relation to how they relate to those from earlier studies, and I attempt to offer explanations for them.

5.3. Fluency in L1 and L2

This section provides a discussion of the findings in Study 1 as they relate to RQ1, RQ1 (a) and RQ1 (b). Starting with the overall differences between L1 and L2 fluency measures observed in the descriptive analysis, I discuss the findings from the correlational analysis and compare these with of those reported in the literature.

5.3.1. Differences in L1 and L2

The descriptive analysis highlighted a number of differences between L1 and L2 fluency measures. The participants showed large differences in the amount of their pausing behaviour in the two languages regarding the frequency of pauses; the figures were higher in English for number of mid-clause pauses (both filled and silent), and in Turkish for number of end-clause pauses (both filled silent). This corroborates the findings of earlier studies (De Jong 2016; Riazantseva, 2001; Tavakoli, 2001; Skehan, 2014) in that while L1 speakers tend to pause more often at clause-boundaries and less so within clauses, L2 speakers tend to exhibit an opposite pattern. As was already discussed in Section 2.4.1.2, L2 learners' mid-clause pausing is assumed to be due to the processing difficulties they encounter in speech production process as their speech is not yet automatic (Kormos, 2006; Skehan, 2014). The results of the t-tests also showed that these differences were statistically significant, which when coupled with the descriptive statistics lends further support to this assumption discussed in the literature.

It was also interesting to see that the figures for several of breakdown and repair measures were very similar in the two languages. The participants performed similarly in L1 and L2 in terms of length of mid-clause and end-clause (both filled and silent), and the number of replacements, reformulations and false starts. The finding for the mean length of silent pauses (both mid and end-clause) contrasts those of De Jong et al. (2015) and Huensch and Tracy-Ventura (2016) both of which reported that duration of silent pauses in L2 were longer than in L1. The different results might be attributed to the way the pauses were identified in these studies and the current one. While both De Jong et al. (2015) and Huensch &Tracy-Ventura (2016) examined pauses within and between *AS-units*, the current study explored pauses within and between *clauses.* Given the research evidence (De Jong 2016; Tavakoli, 2011; Skehan & Foster, 2008) suggesting that L1 and L2 speakers differ in their pause location (i.e. mid and end-clause clauses), making a finer distinction between within and between clauses rather than AS-units could lead to a more detailed and reliable analysis of pausing behaviour.

The descriptives further indicated statistically significant differences between L1 and L2 fluency measures for other repair measures, i.e. number of repetitions and total repair, and speed and composite measures. The figures in English for number of repetitions and total repair were higher, and they were lower for speed measures. This suggests that the participants were overall less fluent in their L2 than in L1. These results are also consistent with those of previous studies (e.g. De Jong et al., 2015; Huensch & Tracy-Ventura, 2016).

It is also noteworthy that overall, the participants' L2 speed (198) was lower than those reported in previous studies (e.g. Mora & Valls-Ferrer, 2012; Tavakoli et al., 2016). The lower L2 articulation rate found in the present study could be explained by the fact that several of the participants belonged with low-proficiency level, i.e. A2. Also, they studied English in an EFL context, with none of them having had previous experience of living or studying abroad for a substantial amount of time. In contrast, the participants in both Mora and Valls-Ferrrer (2012) and Tavakoli et al. (2016) came from higher LP levels, and they either resided in the TL community or were on study-abroad courses. As for L1 articulation rate (378), it was higher than other languages reported by similar studies. For example, L1 (English) articulation rates ranged between 192 to 225 in Tavakoli (2016) and was 280 in Mora and Valls-Ferrer (2012). Given that the present study is one of the first to examine fluency behaviour in the Turkish language as L1, there are no baseline data which the L1 articulation rate[4] could be compared with.

5.4. Correlations

The RQ1 explored whether there was a relationship between L1 and L2 fluency behaviours for lower-proficiency learners (i.e. A2, B1 and B2 at CEFR) in an EFL context. This question was built on the hypothesis that if the relationships are stronger, L2 fluency behaviour would be reflecting the participants' personal speaking styles in L1 more. In answering this question, it was found that L1 and L2 fluency behaviours were, to a certain extent, related to each other for breakdown and repair aspects but not for speed. This partly confirms the findings of previous studies (De Jong et al. 2015, Huensch & Tracy-Ventura, 2016; Peltonen, 2018) in which the correlational analysis also showed correlations for some measures of breakdown and repair fluency, with small to medium levels of strengths. In what follows, I discuss these under separate sections for each fluency aspect.

5.4.1. Breakdown fluency

Regarding the breakdown behaviour, the pauses were examined in their frequency, length, location and character. Four of the breakdown measures correlated significantly between L1 and L2; i.e. number of mid-clause filled pauses (r =. 60), number of end-clause filled pauses (r = .31), number of mid-clause silent pauses (r = .35) and length of end-clause silent pauses (r = .52). With regard to the frequency of filled pauses (both within and between clauses), the current findings are overall consistent with those presented in De Jong et al. (2015), Huensch and Tracy-Ventura (2016) and Peltonen (2018). I should note that these studies examined the total frequency of filled pauses, and

---

[4] The only study examining fluency of Turkish speakers, i.e. De Jong et al. (2015) did not report articulation rate. They reported the mean syllable duration, which is the inverse articulation rate.

unlike the present study, they did not make a distinction between mid and end-clause pauses. Nevertheless, given that the significant positive correlations found for both mid and end-clause pauses in the present study, the findings could arguably be compared with regards to the total frequency of pauses.

As for the silent pauses, it was found that mid-clause pausing behaviour in the two languages correlated significantly. This echoes the finding of Peltonen (2018), in which a significant strong correlation was reported between L1 and L2 fluency (r = .66). However, as was the case with the number of filled pauses, most other studies of this kind (e.g. De Jong et al., 2015; Derwing et al., 2009; Huensch & Tracy-Ventura, 2016) investigated frequency of silent pauses in terms of total instances of silent pausing, which makes is impossible to compare the findings. As was mentioned already, the present study made a distinction between the number of mid-clause and end-clause pauses following from several studies in this area (Kormos, 2006, Riazantseva, 2001; Segalowitz, 2010; Tavakoli, 2011; Tavakoli & Skehan, 2005). These studies have suggested that mid-clause pausing is a key characteristic of L2 speakers when their speech production process has not become automatic yet. The significant correlations found for mid-clause pausing (both filled and silent) here indicate a high degree of association between L1 and L2 mid-clause pausing behaviour. In fact, the highest correlation observed in the correlational analysis was for number of mid-clause filled pauses (r =. 60). Taken together, these findings imply that L2 pausing behaviour within clauses might partly be a function of L1 speaking styles.

As for the length of pauses, the correlational analysis showed significant results only for end-clause silent pauses. This finding is consistent with those of Peltonen (2018) and De Jong et al. (2015) where strong positive correlations between L1 and L2 were reported for this measure (r = .68 and .76, respectively). The results of the present study did not demonstrate any significant correlations for other measures of length, i.e. lengths of mid-clause pauses (both filled and silent) and end-clause filled pauses, implying that length of mid-clause pausing is independent in L1 and L2. Previous research (e.g. De Jong et al., 2013; 2015; Towell et al., 1996) suggested that pause length is not related to L2 proficiency, but it could be linked with some personal styles. Although these studies did not distinguish lengths within and between clauses, it can be suggested that overall, the current findings seem to be in contrast with this suggestion because no correlations were observed for the three measures of length between L1 and L2. It might be that length of pauses could be a characteristics specific to a particular language, as also suggested by Riazantseva (2001) and Derwing et al. (2009), rather than characteristics or a trait of a speaker.

Further, a lack of significant correlation for length of mid-clause silent pauses is somewhat surprising in that the finding parallels that of Peltonen (2018), where no links between L1 and L2 was reported, but differs from that of De Jong et al. (2015), which found a strong correlation for this measure (r = . 65). The contradictory findings might have been caused by two major differences between the studies. First, the participants in each study differed in terms of the amount and/or context of language exposure they had had at the time of the testing. The participants in the current study as well as in Peltonen (2018) studied English in an EFL context and had not lived or studied abroad for any substantial amount of time whereas those in de Jong et al. had lived abroad for an average of 4.5 years. The second one relates to a methodological difference; pause identification in De Jong et al. (2015) was based on within and between AS-units, as opposed to the present study where it was based on within and between clauses. As emphasized several times before, research evidence suggests that mid-clause pausing in L2 speech is an important feature of L2 speech (De Jong, 2016; Tavakoli, 2011; Skehan & Foster, 2008), and therefore making a distinction between mid and end-clause pauses is valuable contribution of the present study. Regarding mid-clause pausing, one tentative implication could be that while mid-clause pausing with regard to the frequency of pauses could, to some extent, be related with L1 speaking styles, with regard to the duration of the pauses it seems to be independent from L1 fluency. However, given the relatively small sample size, these findings must be interpreted with caution.

5.4.1.1. Breakdown fluency across different LP levels

The further investigation of L1-L2 links for breakdown fluency across different LP levels revealed several interesting findings. The study found that the correlations for number of mid-clause pauses was maintained at B1 and B2 levels, but not at A2. It seems that mid-clause filled pauses began to surface in L2 speech as the participants improved in their linguistic knowledge. Previous research has suggested that one reason behind speakers' use of filled pauses might be to facilitate communication (e.g. as a communicative strategy) (Dornyei & Kormos, 1998; Tavakoli et al., 2019). The pattern found for this measure at higher levels of LP seem to support this; it is possible that the speakers' use of filled pauses in their L2 began to resemble their L1 styles (i.e. as a means to facilitate communication) as their improved in their L2 proficiency. However, since the group numbers for LP levels were too small, further research with larger group sizes is recommended on L2 learners' patterns in using filled pauses at different levels as well as in different L1s.

It was also interesting to see that the significant correlations found for mid-clause silent pauses was only maintained at A2 level (r = .62). It seems that L2 mid-clause pausing behaviour for speakers of lower levels was strongly related to L1 styles. Recent studies have reported that the frequency of mid-clause silent pausing distinguishes lower levels of proficiency from higher levels (Skehan et al., 2016; Tavakoli et al., 2019). One possible reason for this could be related to the use of multi-word chunks. Davies (2003), referring to Pawley and Syder (1983), suggests that since L1 speakers rely heavily on the use of multi-word chunks, they can direct their attentional sources to the upcoming message rather than to the formulation of the current one. This mostly results in pausing at clause boundaries. L2 speakers, on the other hand, are more directed to the formulation of the message, and due to their incomplete linguistic knowledge, the speech becomes interrupted with hesitations and pauses within clauses.

Following from the above, it seems that the frequency of mid-clause pausing is informative of L2 proficiency and is therefore likely to decrease in a linear fashion as LP improves. Subsequently, one might expect to see significant (and maybe stronger) correlations at the highest levels only given that L2 speech would become more L1 like. Surprisingly, however, the results failed to indicate any correlations at these levels for mid-clause pausing. One possible explanation could be that although L1 and L2 mid-clause pausing behaviour might be related to some extent, as the correlational analysis showed, L2 proficiency might have had a stronger mediating effect on L2 speech at highest levels (De Jong, 2016) than at lower levels. One could, therefore, argue that it became possible to observe more of the effect of L1 styles at the lowest level, resulting in a correlation between L1 and L2.

When the findings for number of mid-clause filled and silent pauses across LP levels are compared, it would be plausible to suggest that at lower-levels of L2 proficiency, mid-clause pausing occurred mostly in the form of silent pauses (r = .62 at A2 level). However, as the participants expanded their linguistic knowledge, it seems that they started to use filled pauses more often (r = .55 and .84 at B1 and B2 levels, respectively) than they did silent pauses. It is likely that higher-level learners developed communicative strategies to compensate for the processing difficulties which they encountered in the formulation stage. They might have used filled pauses to facilitate communication by filling the silences with non-lexical fillers (e.g. erm, eh, uhm) (Tavakoli et al., 2019). Alternatively, it may be that they might have transferred their L1 pausing patterns to their L2 once they have attained a certain level of proficiency, as also suggested by Riazantseva (2001). Indeed, the high figures for the correlations between L1 and L2 mid-clause filled pauses for higher LP groups seem to confirm this.

Surprisingly, the significant correlation for length of end-clause silent pauses did not show a progressive pattern across different LP levels. The correlations were observed at the lowest, i.e. A2 (r = .53) and the highest level, i.e. B2 (r = .80), with no correlations at B1 level. It is difficult to explain the possible reason for this finding; however, it might be related to the effect of the task type on the performance in L1 and L2 at different levels. As was discussed in Section 2.2.2 , disfluencies occurring in L1 and L2 speech are due to different reasons; i.e. when they have something difficult to conceptualize (Fulcher, 1996) or when they have processing difficulties such as those relating to grammar or lexical retrieval (e.g. De Jong, 2016; Skehan, 2014; Skehan et al., 2016; Tavakoli, 2011), with the former being most typical in L1 and the latter in L2. We might speculate that the participants paused for different reasons while performing the task they were assigned to in their L1 and L2. The participants at the lowest level might have had conceptualization issues in their L1 and processing difficulties in their L2. In such a scenario, they might have produced similar durations of silent pausing in both languages but due to different reasons. This would consequently result in a correlation between the two. In addition, the participants at the highest LP level might have produced similar lengths between clauses in both languages because their L2 speech resembles more to their L1 speech as their linguistic knowledge is more developed. While these could only be some speculations offered to explain the current findings, it seems impossible, at this point, to identify the actual reasons behind the speakers' pausing patterns (e.g. whether due to processing difficulties in the formulation stage or difficulties in conceptualization in the speech production process). An in-depth analysis of the speech samples on pausing measures or retrospective interviews with the participants on their pausing behaviour could be carried out to investigate these. Also, it is also important to note once again that since the group sizes were too small in the present study, these statistical analysis could be repeated with larger group sizes in additional research so that the power of the statistics could be increased.

Finally, no significant correlation between L1 and L2 end-clause filled pauses was observed at any of the levels. It is highly likely that given that the correlation for this measure was already small (r = .30), when the data file was split into three groups and the correlational analysis was repeated, the correlation lost its significance at different levels due to smaller sizes of the groups (n = 15, 15 and 12 for A2,, B1 and B2 groups, respectively).

5.4.2.Repair fluency

The present study found that two repair measures, i.e. number of reformulations and total repair, correlated significantly in L1 and L2 (r = .54 and r = .45, respectively), suggesting that L2 repair behaviour might be, to a certain extent, a function of L1 repair behaviour. As for the other repair measures; i.e. number of repetitions, number of replacements and number of false starts, no significant correlations were observed. With regards to number of repetitions, the findings replicate those presented in Peltonen (2018), which found a weak correlation for this measure, and Huensch and Tracy-Ventura (2016), where number of repetitions correlated only for French group after residency abroad while not for the

Spanish group. As for the results for other repair measures, i.e. number of replacements, number of reformulations and false starts, the results cannot be compared given that most similar studies (e.g. De Jong et al., 2015; Derwing et al., 2009; Huensch & Tracy-Ventura, 2016; Peltonen, 2018) did not examine all of these measures, but usually only one or two repair measures.

The findings seem to indicate that L2 repair behaviour could partly be related with L1 repair behaviour, similar to previous studies which implied that repair aspect of fluency is not connected to L2 proficiency, unlike breakdown and speed aspects (Baker et al., 2014; Cucchiarini et al., 2002; Ginther et al., 2010; Hunter, 2017; Kormos & Denes, 2004) but could be a function of individual differences or traits of speakers. A recent study by Tavakoli et al. (2019) has provided further evidence suggesting that repair behaviour does not distinguish different proficiency levels in a linear fashion. Likewise, in other fluency studies (e.g. Di Silvio et al.; 2016; Huencsh & Tracy-Ventura, 2017), no change was observed over time in LP levels of the participants. In this regard, the current findings lend further support to these studies, in that while repairs could be indicative of underlying monitoring processes in speech production (e.g. reformulating a message), they could also demonstrate L1 personal speaking styles, to some extent.

5.4.2.1. Repair fluency across different LP levels

When the significant correlations for repair measures were further investigated across different LP groups, an interesting picture emerged. The correlations for both number of reformulations and total repair did not show a linear progress across levels; the L1-L2 correlations were only maintained at B1 level only (r = .56 and .60, respectively, for the measures), but not at the lowest, i.e. A2, or the highest, i.e. C1, level. These findings seem to confirm those of Tavakoli et al. (2019), who found a similar pattern in repair behaviour across LP groups. They observed that speakers at A2 level produced the lowest number of repairs and speakers at B1 level the most.

Kormos (2006) argued that a key distinction between L1 and L2 speech production processes is the degree of automaticity in speakers' access to knowledge stores, especially at lower L2 proficiency levels. It might be the case that A2-level learners might not have easily accessed to their declarative knowledge since their speech production has not yet become automatized; subsequently, they may not have engaged in monitoring and repair processes. In contrast, at B1 level, repair processes become activated in both languages (in L2, due to increased linguistic knowledge). Huensch and Tracy-Ventura (2017) explain that as learners improve in their LP, they simultaneously improve in their awareness of making mistakes. This might be possible reason for the findings at B1 level; i.e. the learners might have engaged in monitoring processes highly actively (e.g. by modifying their utterances) due to their increased LP and their awareness in making mistakes. This line of reasoning would mean that at a highest levels, repair processes should be activated even more. Contrary to such an expectation, however, C1-level learners seem to have used repairs in moderation, with results indicating no correlation at this level. One tentative argument might be that L2 repair behaviour was mitigated to different extents by L2 proficiency and L1 personal speaking styles. If it is the case that L2 speakers do not engage in repair processes in a progressive pattern across proficiency levels (Tavakoli et al., 2020), it may also be possible that LP influenced the degree that L1 personal styles surfaced in L2 speech across different levels.

Finally, it can be seen that a similar pattern emerged in the correlations for number of mid-clause filled pauses and number of repairs across LP levels. Repair behaviour has been suggested to be linked to pausing behaviour (e.g. De Jong et al., 2015; Tavakoli et al., 2019); although not explicitly discussed, in De Jong et al. (2015) it was reported that L2 speakers who tended to use more filled pauses also tended to use many repairs. Similarly, Huensch and Tracy-Ventura (2016) observed that Spanish, French and English L1 speakers differed significantly on the use of filled pauses and repair; e.g. Spanish L1 speakers produced fewer filled pauses and corrections than English L1 speakers. As was mentioned, a closer inspection of the current findings illustrates a similar picture for filled pauses and number of repairs at B1 level (and C1 for number of filled pauses), but not at A2 level. Taking these findings together with those reported in the previous studies, two implications could be drawn: 1) the influence of different factors (i.e. L1 styles, cross-linguistic effects and L2 proficiency) might be at work in L1-L2 fluency relationships, to different degrees, 2) repair and breakdown behaviours could be related to each other to some degree. Similar to how L2 learners might use filled pauses as a communitive strategy (e.g. to facilitate communication) as they improve in their LP, they can also use self-repetitions, for example, as ‘a way to buy time’ (Derwing et al., 2004, p.664). Surely, additional research is needed to explore this possible interaction between repair and pausing behaviour.

5.4.3.Speed fluency

As for speed and composite measures, the results did not show any significant correlations between L1 and L2. This finding is contrary to previous studies which did find correlations for speed measures (i.e. De Jong et al, 2015; Derwing et al., 2009; Huensch & Tray-Ventura, 2016). These divergent results can be interpreted in the light of the differences between the participants in the current study and those in other studies. The present study included 15 participants from A2 level (according to CEFR). One common characteristics of lower-level learners is that they often speak at a slower rate and produce more mid-clause pauses, hesitations and interruptions in their speech. This means that their L2 speech is typically slower than L1. On the other hand, participants in most other studies belong with higher proficiency levels; e.g. intermediate to advanced in De Jong et al. (2015) or upper-intermediate in De Jong and Mora (2019). It is likely the case that speaking patterns of lower-proficiency learners in the present study have made it difficult to obtain a significant correlation between L1 and L2 speed fluency. It should also be pointed out that the participants of the present study studied the language in an ELF context, with no prior experience of living/studying abroad or using English for professional purposes in their everyday lives. Given the ample research evidence suggesting that L2 learners benefit from L2 experience particularly in speed fluency (Di Silvio et al.,, 2016; Du, 2013; Huensch & Tracy-Ventura, 2017; Towell et al., 1996), the participants of the present study had arguably fewer opportunities to develop their fluency skills when compared to those in De Jong et al., 2015; Derwing et al., 2009 and Huensch & Tracy-Ventura (2016). By examining the examining fluency patterns of lower-level learners, the current study makes another important contribution to the research in the area.

5.4.4.L1-L2 fluency relationships and other factors

RQ1a and RQ1b set out to explore whether the relationships between L1 and L2 fluency behaviours are mediated by a) an individual learner factor of L2 proficiency and b) an external factor of task structure, respectively. In order to answer these questions, two sets of partial correlations for the significant results obtained through Pearson correlations were carried out, with each of these factors as the controlling variable in each set. Below is provided a discussion of the results for each of the questions.

5.4.4.1. Individual learner variable of L2 proficiency

In order to answer RQ1a, the first set of partial correlations controlling for L2 proficiency was carried out. As was discussed in Section 2.5.1.2, it was hypothesized that L2 proficiency could be a mediating factor in L1-L2 relationships because as L2 learners' proficiencies increase, their automaticity in speech processing does as well, making L2 speech more like L1. This means that the relationships could be overshadowed by the processing difficulties at lower-levels while they could be even stronger at higher-levels. However, this hypothesis was not supported in the current study as the results from the partial correlations controlling for LP did not show any impact of LP on strength of the relationships. Overall, the correlations were maintained regardless of variations in LP. Only for two measures, LP had a very small impact on the fluency relationships; i.e. length of end-clause silent pauses and number of reformulations, with a very small decrease for the former (from .54 to .51) and a very small increase for the latter (from .51 to .54). This suggests that when the impact of LP is controlled for, the strength of the relationship for length of end-clause pauses slightly decreases, suggesting a weaker L1-L2 link. For number of reformulations, however, the strength slightly increases suggesting even a stronger link between L1 and L2. Given that Study 1 was the first study which examined the LP as a mediating factor in L1-L2 fluency relationships, there is no baseline data which the current results could be compared with.

Still, the findings overall could be suggested be in contrast with the suggestions of the research in this area. Several studies have implied that since learners' fluency patterns are likely to change over time as LP increases, the relationships between L1-L2 might be predicted to change as well (Derwing et al., 2009; Huensch & Tracy-Ventura, 2016; Riazantseva, 2001). The present study found that overall L1-L2 fluency relationships persisted regardless of the changes in LP. As such, based on the current results, one can suggest at least with some confidence that L1 and L2 fluency behaviours are positively related to each other and these relationships are not influenced by variations in LP.

5.4.4.2. An external factor of task structure

The significant correlations between L1 and L2 fluency measures were further explored to see if task structure had any impact on the strength of the relationships. The interest of the present study in examining the impact of task structure on L1-L2 fluency links was driven by previous research (Foster & Tavakoli, 2009; Tavakoli & Foster, 2011; Tavakoli & Skehan, 2005); Skehan & Foster, 1999) which reported that task structure affected L2 fluency, but not L1 fluency. This line of research suggested that when a task is comprised of obviously sequenced episodes (i.e. having a tight structure), it becomes less cognitively demanding for the task performer and leads to greater fluency in performance. As such, it was assumed that task structure would be another mediating factor in the strength of the fluency relationships between L1 and L2; performance in a structured task would be associated with stronger relationships between L1 and L2 fluency behaviours.

When controlled for task structure, results from the partial correlations demonstrated that all significant correlations were maintained, suggesting that task structure did not have any impact on the strength of L1-L2 associations. The results of independent samples t-test further confirmed this, indicating no significant differences between the two. This means that the relationships between L1 and L2 fluency measures were similar regardless of whether the task had a tight or a loose structure. Since Study 1 was the only study which examined the mediating effect of task structure on L1-L2 links, again it does not seem possible to compare the findings. When taken together, the results for RQ1a and RQ1b overall suggest that L1 and L2 fluency performance are to some degree linked for breakdown and repair aspects, and these links persist regardless of variations in LP and the structures of the given tasks.

5.5. Predictive power of L1 fluency and L2 proficiency over L2 fluency

The second aim of Study 1 was to investigate whether L2 fluency behaviour can be predicted from L1 fluency behaviour and L2 proficiency. In order to answer this question, multiple regressions for all fluency measures were carried out with L1 measures and LP scores, i.e. the OPT and EIT scores, as the independent variables. The results indicated that eight out of 17 models predicted L2 fluency, with L1 fluency making a significant contribution to six of them, the OPT scores to four and the EIT scores to five. When interpreting the findings, I have followed Plonsky and Ghanbar (2018) who proposed R^2 values in the realm of .20 (or below) and .50 (and above), considered as small and medium, respectively. The models predicting L2 fluency were three breakdown measures; i.e. number of mid-clause filled pauses, number of mid-clause silent pauses, length of end-clause silent pauses (R^2 values of 41%, .24% an .45%, respectively with small to medium values). Below, I discuss how L1 f1uency measures and L2 proficiency scores contributed to these models under two separate sections.

5.5.1. The predictive power of L1 fluency over L2 fluency

The study found that L1 measure made a significant contribution to all significant models for breakdown fluency; i.e. number of mid-clause filled pauses ($p < .001$), number of mid-clause silent pauses ($p < .012$) and length of end-clause silent pauses ($p < .001$). The result for frequency of silent pauses mirrors that of Huensch and Tracy-Ventura (2016), who similarly found that L1 silent pause frequency predicted L2 fluency. A key difference between their study and the current one is that in they measured pauses at AS-unit boundaries instead of measuring at clause boundaries. This is considered a limitation by the literature on fluency (e.g. Peltonen, 2018; Skehan, 2009; 2014), which was addressed in the current study by examining pauses at clause boundaries.

As for the repair models, only two of them were significant; i.e. number of reformulations and total repair (R^2 = 38% and 31%, respectively, both which are small), and L1 fluency contributed significantly to both models ($p < .003$ and $p < .007$, respectively). As was mentioned in Section 3.9.2, there is much less consensus about repair behaviour regarding its link with L2 proficiency with some studies suggesting that this aspect is not indicative of LP (e.g. Bosker et al., 2014; Ginther et al., 2010; Kormos & Denes, 2004). The current findings seem to lend further support to this claim. The pattern observed in the multiple regressions, coupled with those in the correlational analysis (Table 4.5) seem to suggest that L2 repair behaviour is, at least some extent, L1 related and can be predicted from speakers' L1 repair behaviour. These results are overall in contrast with those reported in Huensch and Tracy Ventura (2016) where the models for repair measures did not reach significance. The different results could be interpreted in the light of the differences in repair measures adopted in each of the studies. Huensch and Tracy-Ventura investigated two measures of repair only, i.e. number of repetitions and corrections while five repair measures were used in this study, including total repair which indicates a total of all repair instances. This allowed for the examination of repair fluency from a wider perspective in the current study.

In fact, the findings here also suggest that using the total number of repair instances to gauge L2 repair behaviour could be a better approach in examining repair fluency. This is because all types of repair, i.e. repetitions, replacements, formulations and false starts which have been shown to be monitoring processes that learners engage in speech production (e.g. Kormos, 1998; 2006), are actually closely related to each other. A false start, for example, is likely to lead to reformulation of the content (Hunter, 2017). More importantly, given the complexity of the oral data, researchers often find themselves in situations where they have to make decisions based on their own judgements. Below I try to explain this with an example.

1 | <Henry> (a) the boy <go to> (b) erm <goes to> ski |

2 | <Hen-> (a) the boy <go to> (b) erm <goes to> ski |

In both sentences (b)s are examples of reformulations, as the speaker repeats the word 'go' with some modification i.e. 'goes'. While it is also clear that 2 (a) is a false start, i.e. an utterance which was abandoned before it was completed, the picture is less clear about 1(a). Here, did the speaker replace 'Henry' with 'the boy', or did they simply want to indicate that they were referring to ‘the boy’ by ‘Henry’? If the former case is true, this example would be identified as a ‘replacement’; but if it is not, then there would be no repair instance here. Clearly, the researcher will need to make a decision based on the intonation patterns of the speaker as well. Similar to the example above, a personal observation of the researcher was that Turkish native speakers in the study seemed to repeat the information often intentionally, or to give additional information to intensify the meaning. Such patterns were not treated as repairs here. However, it is important to note that similar issues are likely to occur when analysing speech data, often resulting in disagreements between the coders and more importantly in the results. Future research could consider using the total number of repair instances to examine fluency behaviour.

Further, the model for the only pure measure of speed, i.e. articulation rate, did not reach a significant level, suggesting that L1 speed fluency cannot predict L2 speech. This does not support the findings reported in De Jong et al. (2015) and Huensch and Tracy-Ventura (2016), where mean syllable duration (inverse articulation rate) reached a significant level, with both L1 and L2 proficiency making significant contributions to the model ($p < .001$ and $p < .013$, respectively). In fact, De Jong et al. (2015) found that only for this measure, the equivalent measure which was corrected for L1 fluency predicted LP better. This clearly suggests that L1 articulation rate should predict L2 even at a higher degree. At this point, it is important to recall that L1 and L2 speed measure did not correlate in the correlational analysis either (as was discussed in the previous section). The reasons for the results from the two analyses are likely to be the same. Namely, the majority of the participants were lower-level learners in present study, and therefore, it was likely that their L2 speech was slower than their L1 due to the incomplete linguistic knowledge and less automaticity in using that knowledge (Kormos, 2006). Additionally, these participants had no prior experience of living or studying abroad, compared to those in De Jong et al. (2015). Based on the finding from previous studies (Di Silvio et al., 2016; Huensch & Tracy-Ventura, 2017; Towell et al., 1996) which reported that speech fluency, particularly speed fluency, benefits most from L2 experience, the participants of the present study had arguably fewer opportunities for developing their speech fluency.

Taking the two points together, the current findings imply that at lower levels of LP, L1 fluency does not demonstrate much about one's speed in her L2 speech.

Finally, all the models for composite measures achieved significance; i.e. speech rate (R^2 = 48%), phonation-time ratio (R^2 = 51%) and mean length of run (R^2 = 43%), however, L1 fluency made a modest contribution only to speech rate ($p < .036$), suggesting that L1 could, to a small extent, predict L2 speech rate. In other models for composite measures (i.e. phonation-time ratio and mean length of run), L1 did not predict L2 fluency, suggesting that these L2 composite measures are not related to L1 fluency behaviour but could be a function of L2 proficiency which I will discuss in the next section.

5.5.2. The predictive power of L2 proficiency scores over L2 fluency

Interestingly, L2 proficiency scores contributed significantly to all of the significant models. This is interesting because while the pure speed measure of articulation rate did not reach a significant level, composite measures, which are also speed-related, did. These results further highlight the importance of using both pure and composite measures when investigating speech fluency. Study 1's motivation to examine the role of LP came from previous studies (De Jong, 2016; De Jong et al., 2015; Huensch & Tracy-Ventura, 2016) that questioned whether LP mediated the relationship between L1 and L2 fluency behaviours. The study was also inspired by research that claimed that L2 fluency was one of the most reliable indicators of LP (De Jong et al., 2012a). The results revealed that nearly half of the variance in learners' performance for in each of the composite measures could be explained by L2 proficiency. The variance explained by L2 phonation-time ratio, the model reaching the highest significant level, was 48%, with both OPT and EIT contributing significantly to the model. The variances explained in L2 speech rate and mean length of run were 43 % for both. While L1 fluency contributed only to the speech rate, which was a significant but a modest contribution ($p < .036$), for mean-length of runs, the EIT scores was the only variable which made a significant contribution to the model.

The current findings are very interesting because while the results from the correlational analysis provided in the previous section (4.5) did not indicate any mediating impact of LP on the strength of L1-L2 links, LP scores contributed significantly to several models in the regression analysis. This might be due to the fact that the mediating impact of overall LP (i.e. through the use of a mean score) was investigated in the correlational analysis while the regression analysis sought the predictive power of each LP score separately (i.e. the OPT and the EIT scores) over L2 fluency measures. Although there was a strong correlation between the two sets of scores (r = .51), it is possible that creating a mean score might have led to losing some of the data. On the other hand, entering each score separately into the models in the regression analysis might have helped the explanatory power of each test score surface more. Additionally, the current findings indicate that the EIT scores were a good predictor of all composite measures as well as of number of repairs and mid-clause silent pauses. The OPT scores, on the other hand, was a predictor of mid-clause silent pausing. Two implications could be drawn from these findings. First, procedural knowledge (measured by the EIT) plays an important role in producing smooth and fast speech whereas declarative knowledge (measured by the OPT) may encourage mid-clause silent pausing. Research (Kahng, 2014; Skehan, 2014; Skehan et al., 2016) has reported that mid-clause pausing is linked to processing of the declarative knowledge which occurs in formulation stage (Levelt, 1989), therefore it plausible to suggest that learners use mid-clause pauses as opportunities to access to their declarative knowledge. Second, the findings also highlight the importance of testing LP through more than one means as the two tests, the EIT and the OPT, seem to predict different aspects of performance. Measuring LP from a wider perspective would help us reach more reliable conclusions about the role of LP in fluency performance and fluency relationships between L1 and L2.

5.6. Implications for Study 2

Having discussed the results and observations that have emerged from Study1, I will now briefly discuss the directions for Study 2. While doing so, I mostly reflect on the review of the literature about the measurement of lexical complexity (Section 2.4.2) and attempt to discuss the implications of these for Study 2. A large part of this section is devoted to the implications regarding the lexical analysis since this forms a vital part of the Study 2.

The findings from Study 1 suggested that L1 and L2 fluency behaviours are to some extent related for breakdown and repair aspects but not for speed fluency. It is now necessary to design a new study which extends this investigation in several aspects by exploring L1-L2 fluency relationships 1) in a different learning context, i.e. study-abroad context, which would still allow for a comparison/confirmation of the findings between the two studies, 2) recruiting a higher number of participants, 3) including learners from higher proficiency levels and 4) considering the mediating effects of more individual learner factors such as WMC and LoR. An investigation of L1-L2 fluency links in a further study with these important additions will deepen our knowledge of L1 effects on L2 fluency behaviour and will offer more robust findings about the L1-L2 fluency associations and some additional factors playing a mediating role on these.

Individual variations between L2 learners has been a central area of focus in fluency research. Amongst these, learners' L2 proficiency, WMC or LoR are some of the most important factors whose effects on L2 fluency performance and development are well known. Given the fact that any factors affecting L2 fluency can potentially impact on the strength of L1-L2 fluency relationships, it is essential to further consider the mediating roles of these in order to reach reliable conclusions about the influence of L1 styles on L2 fluency. In this regard, Study 2 aims to explore three individual leaner variables; I.e. L2 proficiency (this time described with three (higher) levels), WMC and LoR.

I have discussed in Section 2.4.2 that a part of Study 2 is also interested in examining lexical complexity in L1 and L2 speech. The review of research suggests that variations in lexical complexity behaviour could be existent at two levels: language level and individual level. Research has shown that lexical performance is sensitive to cross-linguistic effects; with significant differences across different languages under investigation being reported in lexical complexity development (e.g. De Clerq, 2015), and lexical complexity performance (e.g. Daller et al., 2003; Dewaele & Pavlenko, 2003). One possible explanation for such cross-linguistic differences has been typological differences across languages, and in this regard, the most notable factor affecting the performance appear to be how grammatical information is expressed through lexical resources. In Turkish, for example, most function words are embedded in morphemes (e.g. "*masada*", meaning *on the table* in English), therefore, one could argue that Turkish relies less on function words than English does. This implies that function words tend to be higher frequency in English than Turkish, and when language performances in the two are compared in this sense, one may conclude that overall diversity is lower in English. Unfortunately, however, another issue emerging here is that the existing body of research has, to a large extent, focused on English (e.g. Lindqvist et al., 2011) or some other L2s which are not typologically distant from English, e.g. French and Italian in Lindqvist et al. (2011), German in Vyatkina (2012), French and Spanish in De Clerq (2015) or French in Treffers-Daller (2013). This clearly indicates a need for more cross-linguistic research, especially on the examination of lexical complexity for two typologically distant languages, such as Turkish and English in Study 2.

The review of the relevant literature has also revealed that given the multidimensional nature of the lexical complexity construct, different measures are required to tap into each of its sub-dimensions (i.e. lexical diversity, lexical sophistication and lexical density). Existing methods, however, have several shortcomings in drawing a complete picture of the construct (as well as the sub-constructs) since each of the methods has its own limitations (Jarvis, 2013). Therefore, it seems that there is no one single lexical measure that would help us explain the construct fully. Nonetheless, we could suggest, despite the shortcomings discussed before (Section 2.4.2), all measures still tell us some part of the story, with each indicating something about the vocabulary that speakers know and use actively (Milton, 2009). Given the multidimensional nature of the construct and limitations of the existing methods, two useful and informative approaches for researchers would be 1) to examine more than one sub-dimension for a better construct representation, and 2) to apply different measures to the same data rather than relying on a single one.

Related to the measurement of two structurally different languages, one limitation seems to be that there is currently no standard way of measuring lexical complexity in different languages (Daller et al., 2003; De Clercq, 2015; Treffers-Daller, 2013). This could be partly, in some cases mostly, due to the fact that not all existing measures are applicable across languages. For instance, LS measures are based on frequency lists, however, there may be no reliable frequency lists available in every language (in effect, these are available only for a few languages), especially lists based on oral data. This has been why some alternative ways have been suggested such as the use of teacher judgements to determine basic and advanced words used by the participants (Tidball & Treffers-Daller, 2008).

Returning to variations in lexical complexity behaviour, research has also demonstrated that individuals vary in their development of lexical complexity knowledge and performance. Individual variations might surface due to multiple factors, including, but not limited to, L1 background, the closeness of L1 and L2, age or L2 proficiency level (e.g. Dewaele & Pavlenko, 2003), conceptual factors and familiarity with the domain (e.g. Robinson, 2001; Skehan, 1998). To give an example, one could expect a gardener to give more detailed information about kinds of trees or a pilot to talk in length about planes using some specialized vocabulary. This suggests that how much information is presented and/or how detailed and linguistically complex presented information is (i.e. in terms of the linguistic means used such as activities, events, actions) are some of the factors reflected in the lexical performance; e.g. the use of more specialized information could lead to the use of more specific and low-frequency words (ibid.)

On that note, Pallotti (2015) further argues that lexical complexity behaviour is prone to stylistic variation, suggesting that variability is also existent in native speaker data due to the stylistic preferences. What is not clear, however, is whether or to what extent such stylistic variations or personal styles in one's native language are reflected in their L2. In other words, could we argue that a person who talks a lot or use a richer vocabulary in their L1 is also likely to talk a lot and use richer vocabulary in their L2? De Jong and Mora (2019) postulate that individual differences in the oral performance can partly be accounted by differences in the use of skills involved in speech production process (e.g. individual articulatory skills in the articulation stage). More specifically, they suggest non-verbal skills such as non-verbal intelligence (the main individual trait underlying the conceptualization of the pre-verbal message), or executive control skills such as WMC might be lying at the source of individual differences in speech production. Given that all stages (or sub-processes) of speech production in Levelt's model are closely related with each other, such individual non-verbal skills could also be argued to be influential in other stages of the production (e.g. formulation). Additionally, we know that some native speakers have a lower percentage of low-frequency words in their speech; however, this does not necessarily mean that low-frequency words in their L1 are unknown, non-existent or even not easily accessed in their lexicon, but might simply mean that these people have not wanted or *not preferred* to put their knowledge into use (Laufer, 2003) Whether personal preferences or styles are also reflected in L2 speech in terms of lexis used, i.e. whether there is a link between L1 (Turkish) and L2 (English) lexical complexity behaviour, is what Study 2 is interested to find out.

Finally, it is also necessary to note that in Study 2, lexical complexity is represented only through the examination of lexical diversity (LD) aspect in Turkish and English although I acknowledge that the investigation of the construct cannot be reduced to the investigation of LD aspect only. There are a few reasons for not being able to explore other aspects, i.e. lexical sophistication (LS) and lexical density, in Study 2. With regard to the LS aspect, 1) to the best of my knowledge, no reliable frequency list is available in Turkish to measure Turkish LS with the exception of a few such as Turkish National Corpus (Aksan et al., 2012), which is quite limited and mainly based on written data), and 2) for reasons of practicality and time, alternative ways such as the use of teacher judgements to determine the basic and advanced words in participants' speech production (Daller et al., 2003; Tidball & Treffers-Daller, 2008) could not be sought after. As for the density aspect, the two languages under examination may not be comparable due to the structural differences between them; Turkish a morphologically dense language in that function words are embedded in bound morphemes, thus the number of function words in Turkish and English are likely to differ to a large extent. For these reasons, only LD measures provided a practical way in the study to examine the two languages in terms of the lexical complexity. I will explain these further in Section 6.10.2).

5.7. Summary of the chapter

In this chapter, I have discussed the findings of Study 1 as they relate to the two RQs (with sub-RQs) of the study under two main sections. In the first section, I explored the relationship between L1 and L2 fluency behaviour and whether this relationship is mediated by variations in an individual learner variable of L2 proficiency, and an external factor of task structure. I linked the findings emerging from simple and partial correlations to the relevant literature in this area. I suggested that L1 and L2 fluency behaviours are linked to each other, at least to some extent, breakdown and repair aspects but not for speed fluency. Based on the strengths of the correlations between fluency measures in both languages, I argued that particularly breakdown and repair fluency in L2 might reflect L1 speaking styles. I also discussed that these relationships overall appear to be independent of variations in LP, and task structure.

The second section focused on the predictive power of L1 fluency and L2 proficiency, represented through the OPT and the EIT scores. I discussed the findings emerging from multiple regressions and explained that while a number of number models reached significance, L1 fluency and L2 proficiency scores predicted contributed differently to the models. The overall picture suggested that L1 fluency contributed significantly to a number of breakdown and repair measures. The EIT and the OPT scores, on the other hand, predicted different aspects of fluency performance: the former predicted speed-related fluency measures while the latter predicted mid-clause pausing. I argued that the choice of test is important as the two tests appear to test different knowledge types (i.e. the OPT assumed to test declarative knowledge and the EIT assumed to test procedural knowledge).

CHAPTER 6. METHODOLOGY: STUDY 2

6.1. Introduction

In this chapter I present the methodology of Study 2 in nine sections. I begin with the aims of Study 2, the research questions and the corresponding hypotheses that guided the study. Then, I move on to describe the research design with the variables of the study and provide a detailed account of the instruments used with their justifications. A detailed discussion of the pre-pilot and the pilot study is also presented, and the procedures of Study 2 are explained in-depth, including the setting of Study 2 and how the data was collected. Next, I provide the steps that were followed in the data analysis, covering data transcriptions, the measures that were adopted, the data coding and inter-reliability issues. Finally, I conclude the chapter with a summary.

6.2. Aims of the study

While Study 1 aimed to investigate whether there was a relationship between L1 fluency behaviour and L2 fluency behaviour and was therefore exploratory in terms of the wide range of fluency measures employed, Study 2 is extending the findings of Study 1 in examining fluency and exploratory in investigating lexical complexity. As discussed in the previous chapter (Chapter 4), the motivation for Study 2 comes from both the existing evidence that L1 and L2 fluency behaviours are linked to each other, at least to some extent, and the findings of Study 1, which partially confirmed previous studies' results and demonstrated correlations between some of the fluency measures used. The current study aims to expand our knowledge of this link between L1 and L2 oral performance with regard to fluency in several directions. Firstly, it examines the relationship in a study-abroad context as opposed to Study 1, which was carried out in an EFL context. It is hoped that Study 2 would develop our understanding of cross-linguistic effects on L2 speakers' fluency behaviour by incorporating higher L2 proficiency levels and extending the context from an EFL to a study-abroad one.

Secondly, similar to Study 1, Study 2 aims to investigate the role of L2 proficiency level in mediating the relationships between oral performances in both languages. LP, an individual factor, has been shown to be directly related to L2 development and acquisition in L2 studies (Hulstijn, 2012; 2015). Research suggests that although L1 and L2 speech production essentially rely on similar processes, the degree of automaticity in the two is different (Kormos, 2006). In L1 speech production, parallel and automatic processing mechanisms, which require no or less conscious attention for the speaker, are at work, which helps make L1 speech more fluent. In contrast, in L2 speech production, the processing is not typically automatized; the access and retrieval of words and rules from the mental lexicon and the monitoring of the language require more conscious attention for the L2 speaker. This is especially the case for speakers of lower proficiency levels due to their incompetence in the linguistic knowledge. For such L2 speakers, the controlled processing is likely to result in signs of disfluencies such as slower speech or repeated pauses. However, as LP improves, L2 speech is less affected by the processing difficulties and becomes more automatized. The implication of these for the current study is that the mediating effect of LP on the relationships between L1 and L2 fluency behaviour might change across different proficiency levels; i.e. at lower proficiency levels, the processing difficulties encountered in L2 speech production might result in a larger impact of LP on the strength of the relationships than at higher proficiency levels. This was partially tested with L2 speakers of lower proficiency levels in Study 1 and the findings demonstrated that the LP did not mediate the relationships between L1 and L2 fluency behaviour; however Study 2 aims at investigating the fluency behaviour in L1 and L2 at higher levels of LP, and therefore it is of particular importance to the current study to examine LP.

As such, the participants in the current study belong to higher levels of proficiency (i.e. B1, B2 and C1) compared to those in Study 1, where the participants were from lower proficiency levels (i.e. A2, B1 and B2). In this regard, similar to Study 1, Study 2 is also one of the few studies where participants belong to three different proficiency levels rather than only two groups, which is most common in previous studies. Given the importance of LP in fluency studies, it is hoped that working with three proficiency levels rather than two would help us examine the relationship between L1 and L2 fluency behaviour from a broader perspective and lead us towards a better understanding of L2 oral performance.

In addition to the role of the different proficiency levels on the relationship between L1 and L2 fluency behaviour, Study 2 further aims to understand whether working memory capacity (WMC), as another individual factor, and the amount of time that the participants have spent abroad, i.e. LoR, mediate the relationships between L1 and L2 fluency behaviour. As already discussed in Chapter 4, the motivation for incorporating WMC to the current study as another independent variable comes from the existing research evidence which suggests that individual differences in WMC are influential in L2 development and L2 performance (Wright, 2013, 2015). As mentioned before, speech production has been shown to be a complex process which involves a set of mechanisms that operate simultaneously and that require attentional and linguistic resources, especially for L2 speakers (Kormos, 2006). WMC is suggested to support this complex process by managing attentional resources available. For example, in the conceptualization stage of the speech production, conceptual chunks are held in the memory while a series of other cognitive processes take place; e.g. lexical access, organization of morphological or phonological forms, monitoring the language and maintaining the fluency. Given that a continued allocation of attention is needed for these ongoing processes and maintaining the information in the memory (Skehan, 2015), it is assumed that higher WMC allows for more attentional resources to be freed up and dedicated to complex management of verbal input and output. Therefore, it is suggested that WMC could aid speech production and fluency (Wright, 2015), especially at lower proficiency levels in L2 speech production. In relation to this, it can be hypothesized that WMC, as a cognitive ability factor, might be at stake in mediating the relationship between L1 and L2 relationship by overshadowing the performance at lower levels of L2 proficiency, which is what the current study is interested in.

Finally, at the time of the study, the participants of the current study were residing abroad for varying amounts of time, therefore the length of their residence (LoR) in the UK was added as another independent variable. Several studies have shown that when the context of learning a language is compared, i.e. at-home instruction and study-abroad, those learners in a study-abroad context make significantly more improvements in their oral abilities than their counterparts receiving traditional instruction, i.e. at-home instruction (e.g. Freed et al., 2004; Serrano et al., 2011). This line of research further indicated that the amount of time spent abroad contributes greatly towards gains in oral performance, particularly in speech fluency (Du, 2013; Huensch &Tracy-Ventura, 2017; Kim et al., 2015, Mora &Valls-Ferrer, 2012). In other words, the more learners are exposed to language abroad, the more fluent they become in their L2 speech. Since the current study is carried out in a study-abroad context and the length of the time that the participants have spent in the UK varies, the participants' L2 oral performance is likely to be affected by this factor (i.e. LoR). Given this, it would not be wrong to assume that LoR may exert an impact on the strength of the relationship between L1 and L2 performance as well. Thus, it is deemed necessary to examine whether LoR, as another individual variable, mediates the relationship between L1 and L2 fluency behaviour.

In addition to examining fluency of performance, Study 2 also aims to explore the lexical complexity of the oral performance in the two languages the participants speak. Examining lexical complexity in this study will help us 1) understand whether lexical complexity in L1 and L2 performances are linked to each other, and 2) understand whether the two aspects of oral performance, i.e. fluency and lexical complexity, are linked to one another. The inclusion of lexical complexity in the current study is justified in a number of ways (please see Sections 2.4.2 and 5.6 for a detailed discussion). First, one of the main assumptions adopted in the study concerns the possible link between fluency and lexical complexity; the hypothesis is that they could affect each other in the formulation stage of Levelt's speech production model. In other words, it is assumed that the processing difficulties encountered in the formulation stage (especially lexical retrieval problems) are likely to result in disfluencies in the overt speech. Previous studies reported that lexical retrieval difficulties lead to more frequent pauses between utterances, no matter what language, i.e. L1 or L2, is spoken (e.g. De Jong, 2016). However, these studies did not investigate Turkish as L1 and English as L2, and therefore, we need more cross-linguistic studies examining the link between fluency and lexical complexity in the oral performance of a given language.

Secondly, these two aspects, namely fluency and lexical complexity, have been commonly reported as stronger indicators of L2 proficiency (e.g. Iwashita et al., 2008; Malvern & Richards, 2002; Revesz et al., 2016; Yu, 2010). Finally, recent studies suggest that the lexis used in one language could affect the lexis used in another or second language; this could be due to the typological differences or similarities, for instance. However, cross-linguistic studies in this regard are very scarce, and the existing research has largely focused on either English or other L2s that are structurally similar to each other. Therefore, there is a clear need for more cross-linguistic studies with different languages (Turkish and English in this case). In summary, in addition to exploring fluency behaviour, the current study intends to explore the relationship between L1 and L2 lexical complexity and between fluency and lexical complexity within both languages (i.e. between L1 fluency and L1 lexical complexity, and between L2 fluency and L2 lexical complexity).

6.3. Research questions and hypotheses

Study 2 has been guided by the following research questions and the hypotheses:

RQ1) To what extent are L1 Turkish and L2 English fluency behaviours related among higher-proficiency learners in a study-abroad context?

H1) Following from the findings of Study 1, L1 and L2 fluency behaviours, with regard to breakdown and repair fluency aspects, will be related to each other at least to some extent.

1a) Does level of L2 proficiency mediate the relationship between L1 and L2 fluency behaviours among high-proficiency learners in a study-abroad context?

H1a) Following from the findings of Study 1, L2 proficiency level will not mediate L1-L2 fluency relationships.

1b) Does working memory capacity (WMC) mediate the relationship between L1 and L2 fluency behaviours among high-proficiency learners in a study-abroad context?

H1b) WMC will mediate the L1 and L2 fluency relationships by promoting L2 fluency for learners with higher WMC.

1c) Does length of residence (LoR) mediate the relationship between L1 and L2 fluency behaviours among high-proficiency learners in a study-abroad context?

H1c) LoR will mediate the L1-L2 fluency relationships by promoting L2 fluency for learners with longer LoR.

RQ2) To what extent are L1 Turkish and L2 English lexical complexity behaviours, represented through LD scores, related among higher-proficiency learners in a study-abroad context?

H2) L1 and L2 lexical complexity behaviours will be related to each other at least to some extent.

RQ3) To what extent are fluency and lexical complexity aspects of second language oral performance for both languages (Turkish and English) related among higher-proficiency learners in a study-abroad context?

H3) Fluency and lexical complexity aspects in each language, I.e. Turkish and English, will be related to each other at least to some extent.

6.4. Design

Similar to Study 1, Study 2 adopted a quantitative approach to researching this topic by collecting samples of the speakers' performances, coding them into numerical values and analysing them through statistical analyses, (Dörnyei, 2007). In the study, the analysis of fluency performance of (e.g. number of pauses, lengths of pauses or d measure) or test performances (e.g. scoring of language tests or working memory tests) require the use of numerical data (Cohen et al., 2018; Lowie & Seton, 2013; Mackey & Gass, 2011).

A mixed within-between participants factorial design was employed to investigate the effects of different variables on participants' performance in the two languages. The study had three independent variables with LP level being a between-participant variable with three levels (i.e. B1, B2 and C1) and with WMC and Lora being continuous variables. Oral performances on two different tasks in both languages (i.e. L1 and L2) are within-subject variables of the study and these are represented through fluency and lexical measures. The design of Study 2 and the variables are presented in Table 6.1below.

Table 6.1. Study design and the variables of Study 2

Study Design	Independent variables	Dependent variables
Mixed factorial design within-between participants (N= 60)	▪ L2 proficiency level with three levels: - B1, B2 and C1 ▪ WMC (used as a continuous variable) ▪ LoR (used as a continuous variable)	▪ L1 speech performance (represented through fluency and lexical complexity measures) ▪ L2 speech performance (represented through fluency and lexical complexity measures)

6.5. Instruments

6.5.1. Background questionnaire

The background questionnaire used in this study is an extended version of the one used in Study 1. Given the change of the focus in this study, i.e. the inclusion of LoR as another independent variable, the new questionnaire was modified to collect information about the amount of time that the participants had spent in the UK. The questionnaire comprised of two sections; while the first section enquired demographic information about the participants (e.g. age, gender, level of education, etc.), the second section elicited information about their language profile (e.g. number of years of language instruction they had received before coming to the UK). The amount of time that participants had spent in UK (i.e. LoR) was added as a new question to the first section (Appendix K).

For the same reasons explained in Study 1 (Section 3.5.1), the purpose of employing this questionnaire was to control for learner individual variables such as age and educational background to ensure that there was not a large variation amongst the participants that could affect the results. For example, those whose first language was not Turkish, despite the fact that they had been brought up in Turkey and they knew Turkish very well, were excluded from the study. Similarly, those who had been living in the UK for more than 5 years or professional purposes (e.g. for work), or those who had been married to people of other nationals and had been using English continuously at home (regardless of that they were also students at the university at the time of the study), were not included. It was assumed that such differences could affect the homogeneity of the sample and the findings. Therefore, the participant pool consisted of Turkish native speakers who 1) were aged between 22-35, 2) had completed at least a bachelor degree in Turkey, 3) were studying at a post graduate degree in any subject in the UK, 4) had not been living in the UK for more than 5 years and 5) had not been using English continuously whether for professional or family reasons at the time of the study. The background questionnaire was given to the participants before the start of data collection to determine, as early as possible, whether they were eligible for the study or not.

6.5.2.Proficiency tests

6.5.2.1.Oxford placement test and elicited imitation task

In Study 1, all the participants took OPT (Alan, 2004) and EIT (Ortega et al., 2002). For practical reasons, only the grammar part of the OPT was employed in the current study, and this part was mainly a pen-and-paper test lacking an oral component. Although the grammar component of the OPT has been considered a reliable tool and has been used widely in several studies, a sole reliance on the results of this test to determine L2 proficiency levels of the participants was considered a potential limitation for the study (Awwad & Tavakoli, 2019) since the primary focus of the current study was on oral performance. Hence, it was deemed necessary to complement the results from the OPT with another valid tool, i.e. EIT, and use a combined score of the two tests to identify L2 proficiency levels. Thus, the EIT was included as another measure of L2 proficiency; i.e. to serve as an oral component in the measurement of L2 proficiency.

Another reason for using a combination of these two proficiency tests came from the existing research evidence which suggests that the OPT is a standardized written test of declarative knowledge test (Elder & Ellis, 2009) while the EIT is a measure of procedural knowledge test (Gaillard & Tremblay, 2016; Wu & Ortega, 2013; Yan, et al., 2016). I fully acknowledge that neither of the tests can be considered as a pure measure of declarative or procedural knowledge. That is to say, it is difficult to determine whether test takers rely solely on their declarative knowledge or use some of their procedural knowledge as well when taking an OPT, or similarly one cannot surely assert that test takers use their procedural knowledge only when listening and repeating sentences in English in an EIT (Suzuki & De Keyser, 2019). However, the assumption adopted in this study was that using two tests in combination as suggested by research (e.g. Awwad & Tavakoli, 2019), rather than relying solely on one grammar test (i.e. the OPT), would tap into different aspects of L2 proficiency; i.e. declarative and procedural linguistic knowledge (De Keyser, 2003, 2009; Hulstijn, 2005) and would yield a more accurate assessment of this construct (please also see Section 2.5.1.1 for a detailed discussion of these linguistic knowledge types).

As such, in Study 2, these two proficiency tests were employed to measure L2 proficiency and to group participants into different levels. A detailed explanation of these is already provided in Section 4.1. Therefore, it would be sufficient to note here that the OPT is a multiple choice written test (a pen-paper version) comprising a range of grammar and lexical items; test takers are presented three options for each question and are asked to select the correct answer. The EIT, on the other hand, is essentially a listening-to-speaking task where test takers are required to listen to a set of 30 sentences of varying length (7-19 syllables) one by one and to repeat each sentence as accurately as possible. As already mentioned, the latter was employed to complement the OPT's limitation in that the OPT does not include any language production.

At this point, it is also worth mentioning that no proficiency test was used to measure proficiency in Turkish (as was the case in Study 1), because all the participants of the study were assumed to have similar educational background as they all had completed at least a bachelor degree in their home country and were studying, at the time of the study, at their post-graduate degrees (i.e. either master's or PhD degrees) in the UK.

6.5.3.Oral narrative tasks

As discussed earlier, oral narrative tasks have been frequently used in SLA research, especially when assessing second language performance (e.g. Iwashita et al., 2008; Robinson, 2001) as well as in language testing contexts (e.g. Test of Spoken English). Similar to Study 1, the choice of oral narrative tasks in Study 2 was partly motivated by 1) practical reasons ; e.g. the ease of administration and being less time-consuming , and 2) by the fact that it had proved too difficult to find or design two identical tasks in other task types (e.g. silent videos). In Study 1, the two oral narrative tasks used to elicit speech samples from the participants (i.e. Henry tasks, see Appendix C) were different in their structure. In Study 2, I opted to replace Henry tasks with other narratives for two reasons. First, the Study 1's participants did not produce a lot of speech, with the majority of speech samples lasting less than 60 seconds, which suggested that the tasks were not suitable for collecting longer samples of the participants' speech. This could be attributed to the nature of the tasks in that each task comprised of eight picture stimuli and was relatively simple and easy to tell. Secondly, task structure was an independent variable in Study 1, and therefore, what primarily made these two tasks different was their structure, namely how well connected the prompts were to each other. However, since the findings did not indicate any effect of task structure on the relationship between fluency behaviours in L1 and L2, this variable was not included in the current study, and therefore, there was no longer a need for the two tasks to be structurally different. Given that Study 1 intends to extend the findings of Study 2 in examining fluency links between L1 and L2 fluency behaviours, one might argue that changing the tasks used might affect the comparability of the findings. Here, it is important to remind the reader that although Study 1 and Study 2 used two different tasks (i.e. Henry tasks and Tom & Jerry tasks) to elicit speech samples from the participants, both tasks were essentially of the same task type, i.e. *oral narrative tasks*, which were monologues and required the task performer to narrate a story by looking at a set of pictures. Therefore, although participants would be expected to use different lexical items due to the different stories involved, it was assumed that speech performances in both studies could still be compared as far as fluency was concerned.

For these reasons, two new oral narrative tasks were designed to be employed in the current study. In order to create two picture stories suited to the purposes of the study, i.e. to elicit speech samples in both languages, once again a number of sources were searched, including EFL sources such as course books and online sources such as videos or cartoon clips. Following from the research suggesting that more familiar tasks lead to more fluent speech (Foster & Skehan, 1996; Skehan & Foster, 1997), the new tasks were created from scratch using Tom and Jerry Cartoons. The reason for choosing Tom and Jerry cartoons was that most people were presumably familiar with these cartoons. Recommendations from De Jong and Vercellotti (2016) (discussed in Section 3.5.3) were followed to design the tasks; every effort was made to create two tasks which were identical to each other in terms of the number of characters and prompts. In addition to this, other factors were taken into consideration; e.g. storyline complexity (Tavakoli & Foster, 2011), amount of intentional reasoning (Awwad, et al., 2017) and the amount of contextual support (Revesz, 2009; 2014) as well as the clarity of the pictures and the story, appropriateness for the task-performers, the vocabulary and linguistic structures required to narrate the stories and the length of the stories.

Initially, three tasks were created. Screen shots were taken from three cartoon series and these were sequenced in a meaningful order to set up a story for each series. In order to ensure that a) the tasks were well suited to the purposes of the study, and b) the most similar two out of the three could be selected, teachers' and PhD researchers' views were sought. For this purpose, a questionnaire was designed, which, together with the three tasks, was given to 6 EFL teachers and L2 researchers who were experts in the field (please see Appendix L for a copy of the questionnaire).

The three tasks were named appropriately, based on the theme of the story that developed in each one (i.e. Task A- Quiet Please, Task B- Best Friend, and Task C- Jerry and the Goldfish). There were three characters in each of the tasks and the picture stimuli were sequenced in a way that each task had a clear beginning, middle and an end. Also, task instructions were provided at the beginning of each task for the participant. These tasks were followed by the questionnaire, which contained three open-ended questions and started with an explanation for the teachers and researchers about the purpose of the questionnaire and the tasks. The open-ended questions sought the experts opinions' about the tasks' comparability regarding a) difficulty for the participant, b) performing the task and linguistic demands (e.g. structure, vocabulary needed) and c) eliciting language samples (vocabulary, range of language expressions, etc.). After these questions, they were asked to rate each task in terms of the difficulty (i.e. easy to understand and to tell) and of being interesting to the prospective task takers on a 6-point scale (with 1 being very easy/interesting and 6 being very difficult/boring).

The motivation behind giving this questionnaire to the experts in the field was to seek help from them in order to make an informed decision on the selection of the tasks. Therefore, the responses were analysed qualitatively and holistically. The responses from both EFL teachers and researchers overall suggested that the first two tasks (i.e. A and B) were more comparable in terms of the language to be elicited and the difficulty for the participants. While not all of the experts found these two tasks the most interesting ones for the participants, I still opted to employ them for the current study because the first two criteria on the rating scale, i.e. the comparability in terms of the linguistic demands and the difficulty, were more important for the purposes of the study than the third criterion, i.e. being of interest to the participants. Consequently, two of the Tom and Jerry tasks (Appendix M), i.e. Task A- Quiet Please and Task-B Best Friend, were selected to be used to elicit speech samples from the participants.

Both oral narrative tasks were made up of 12 pictures in total, and the order of the pictures were mostly sequenced with little or no possibility of changing the order without compromising the main theme of the story. They both had clear storylines without complications (e.g. flashbacks). Each picture stimulus was numbered, and, as was mentioned above, the instructions and the names of the characters were provided at the beginning of each task. The characters in both were the same; there was a dog, a cat (i.e. Tom) and a mouse (i.e. Jerry). In Task A (i.e. Quiet please), Tom was chasing Jerry, first with a pan then with a riffle in his hand, while the dog was trying to sleep and getting disturbed. In Task B (i.e. Best Friend), Tom was again chasing up Jerry to eat her, but Jerry got saved by the dog since she and the dog had already become friends. Both tasks were piloted before the data collection started. The details of the pilot study are provided in Section 6.6.).

6.5.4.Working memory capacity tests

Two types of WMC tests have been used commonly by researchers; i.e. short-term memory tests or simple span tests, which measures only the storage element, and complex WM tests or complex span tests, which tap into both storage and processing elements of the memory (Baddeley et al., 2009) (please see Section 2.5.2.2 for a detailed discussion). Wen (2015) suggests that a simple span task could be particularly useful when working with 'less-educated (participants) or those with low levels of literacy' (p. 35), and a complex span task could be added as another measure when measuring relatively high-proficiency L2 learners. Regarding these tests, Gathercole (2007) further argues that one single task cannot estimate WMC, suggesting that there is not a pure single measure of this capacity. Following from these, in the current study two WMC tests were employed 1) to represent each of these elements of the memory (i.e. storage and processing) and consequently, 2) to capture a better picture of this complex system. Using a combination of two tests would also lead to more reliable and valid results, since it would increase the representability of the WMC. Serial Non-word Recognition Test (SNRT) (O'Brien et al., 2006, 2007) was chosen as a simple span task whereas Backward Digit Span Test (BWDS) (Awwad & Tavakoli, 2019) was adopted as a complex span task. In the following sections, a brief discussion of each of these tests is presented.

6.5.4.1. Serial non-word recognition test

SNRT has been used as measure of phonological memory, which has been established by the research to have a role in language development both in L1 and L2 (French & O'Brien, 2008; O'Brien et al., 2006, 2007). The role of PM, particularly in the development of speech production, has also been examined by several studies (Miyake & Friedman, 2012). This body of research suggests that PM exerts an impact on the acquisition of chunks of word forms and phonological forms influencing developmental domains in SLA such as acquisition of grammar, formulas and vocabulary. Related to this, PM is likely to contribute to the development of L2 narrative skills and speech production, especially in the early stages of L2 proficiency and to the development of grammatical competence at later stages (O'Brien et al, 2006; Wen, 2015).

PM (or WM as is referred henceforth) in the current study was measured using a SNRT. This task was adapted from O'Brien et al. (2006). The advantages of the SNRT were the 'low spoken output demands' (Gathercole, Pickering, Hall, & Peaker, 2001) and that it was 'less susceptible to long-term effects' (O'Brien et al., 2006, p. 380) (e.g. lexicality or language of testing). Therefore, the SNRT was arguably considered a pure measure of phonological storage (Baddeley, 2003). This task required the participants to listen to several lists of non-words (i.e. words that are not meaningful and do not exist) twice and decide whether the non-words in the second set occur in the same order as the first one. The non-words used conformed to the English phonotactic rules (i.e. one-word syllable consonant-vowel-consonant stimuli). The procedures followed in the implementation of the SNRT are explained in the next chapter (Section 6.8).

6.5.4.2. Backward digit span test

Another WMC test used in the current study was BWDS, which has been recognized and used as a measure of complex WMC in a number of studies (Kormos & Safar, 2008). Complex WMC involves both central executive function, which is suggested to control attentional resources, and phonological loop (Wen & Li, 2019). Therefore, complex tests measuring this capacity such as BWDS are assumed to measure both processing and storage features of the memory (Gilabert & Munoz, 2010; Kormos & Safar, 2008; Kormos & Trebits, 2011; Mitchell et al., 2015). Furthermore, a major advantage of such digit span tasks as BWDS is that they are language independent since no language use is required, which also lessens language comprehension demands. This might indicate that digit tasks are particularly useful in that any impact of test takers' language proficiency on the test performance could be minimised (Harrington & Sawyer, 1992; Wright, 2010) For this reason, BWDS was deemed appropriate for the purpose of the current study; namely a) to measure both storage and processing features of the memory and b) to complement SNRT to have a more representative picture of WMC.

An English version of BWDS test was used in this study, which was adopted from Awwad (2017). It is argued by some researchers that WMC testing should be done in L1 to remove the impact of LP as a confound (e.g. Gilabert & Munoz, 2010). A WMC test in L1 was not used since recent studies have reported that WMC scores in L1 and L2 correlated significantly with each other (e.g. Awwad &Tavakoli, 2019). This suggests that both tests, i.e. BWDS in L1 or L2, perform equally well since test takers do not perform significantly different on each one. The test was comprised of sets of varying number of digits (i.e. 3-9 digits). The procedures followed in this task are detailed in the next chapter (Section 3.8.2).

6.6. Pilot study

Prior to the commence of Study 2, a pilot study was designed to find out a) whether the new oral narrative tasks created for Study 2, i.e. Tom and Jerry tasks, and WMC tests were in line with the aims of the study, and b) whether any amendments were needed. The proficiency tests (i.e. OPT and EIT) were the same as the ones used in Study 1; however, in Study 1, the OPT was given to all the participants in a classroom and the majority of them came from lower levels of L2 proficiency. In Study 2, on the other hand, all the data collection procedures would take place in individual sessions, and for this reason, a better estimate of the time required for the whole procedure per individual was needed. Thus, although already used in Study 1, the proficiency tests were also added to the pilot study.

Seven native speakers of Turkish were recruited in the pilot study; with four of them being males and three females. Four of these participants took part in the pre-piloting of oral narrative tasks and the remaining three in the piloting of all materials (Table 6.2). The participants were aged between 27 and 35 and had been residing in the UK for more than three years. At the time of the pre-pilot and the pilot study, they had either completed their master's degrees or had been studying in their PhD degrees at varying subjects in the UK. Thus, their L2 proficiency levels ranged between B2 and C1. These participants were intentionally selected for the pilot study since their profile, in terms of the age, first and second language background and educational level, resembled closely to the targeted participants for the main study (i.e. Study 2).

First of all, narrative tasks and WMC tests were pre-piloted with four participants in one individual session since these tools were totally new additions to Study 2. The participants were given the two oral narrative tasks; with one to be performed in Turkish and the other in English, and the order of the starting language was counterbalanced. These were followed by WMC tests, i.e. BWDS (Appendix N) and SNRT (Appendix O). Their performance on WMC tests was tracked on a sheet manually and was audio-recorded. The pre-pilot study took place in a quiet room at researcher’s house for reasons of practicality.

The results of the pre-pilot study revealed that two of the participants became confused about the names of the characters. In other words, they were not sure whether they were required to use the characters’ names, or they could simply refer to them as cat, dog and mouse. This hesitation was reflected in their performance; it was observed that they were continuously attempting to correct themselves about their use of names, which led to frequent pauses and hesitations. This was an important observation since it was signalling that the actual oral performance of the participants in the main study, especially in fluency aspect, was likely to be affected as well. For this reason, this point was later made clear to other three participants in the pilot study and to all participants in the main study by including the names of the characters in the instructions. They were also told that they could refer to the characters in the way they wished to so that they could be free of concerns over this issue and feel as comfortable as possible when narrating the stories.

As for the WMC tests, the participants reported that before the start of the actual test in SNRT, they were not sure if the words they were going to listen were going to be completely meaningless, and thus at the start of the test, they made an effort to understand the words. Although this issue was addressed by providing clearer instructions for this task in the pilot study and in the main study, it was not seen as a big problem because SNRT started with a small practice and the test format would all be clear in this small section. Regarding the BWDS test, the participants reported that the task was clear, and they did not have any problems in understanding or performing it.

Once the pre-pilot study was completed, all the materials to be used in Study 2 were piloted with the remaining three participants (Table 6.2). In the pilot study, two individual sessions were arranged for each participant in a quiet office. In the first session, each individual was given an information sheet for them to read and a consent form to sign, before they sat the OPT. In the second session, which took place a few days later, they performed the oral narrative tasks (one task to be told in each language), the EIT and the WMC tests, starting with BWDS and finishing with SNRT. Their performances were recorded on a digital audio recorder during this second session in order to detect any technical issues and to be able to make a better judgement about the quality of the recording. At the end of the piloting, each participant was asked to raise any concerns or issues regarding the clarity of the materials and their performance.

Table 6.2. Tools, number of participants and sessions in the pre-pilot and pilot study of Study 2

(N=7)	Number of sessions	Tools piloted
Pre-pilot study (n=4)	One individual session	• Two oral narrative tasks - One was perfomed in Turkish and the other in English - The order of the starting language was counterbalanced • The WMC tests, i.e. the SNRT and the BWDS

Pilot study **(n=3)**	Two individual sessions	Session one (an approximate of 50 minutes including instruction time):: • The OPT Session two (an approaximante of 30 minutes including instruction time): • The EIT • The two oral narrative tasks (Tom and Jerry) - One was performed in Turkish and the other in English - The order of the starting language was counterbalanced • The WMC tests, i.e. the SNRT and the BWDS

The results of the pilot study revealed that the instructions, the tasks and the tests were clear and well-understood. The participants reported that the materials were appropriate, and they did not have any difficulties in understanding the procedures. Only one of the participants expressed his concerns about his familiarity with the oral narrative tasks; he was worried that he was already familiar with these cartoon series and asked if this was acceptable for the study. Since task familiarity was one of the selections and design criteria of the narrative tasks, no action was required for this issue.

6.7. Ethical considerations

Similar to Study 1, the current study followed the Ethics Guidance of Reading University. Before the commence of the study, it was ensured that the study and the data collection procedures adhered to the University's Ethics Guidance. In order to obtain ethical clearance, I submitted an application for the approval to the Ethics Committee of the School, which was granted approval (Appendix E). During the data collection, the confidentiality and privacy of the data were ensured by the researcher. The participants were informed about their right to withdraw from the study at any time they wished to, and this point was also made clear in the information sheet. They were further told that at no point at the study their names would be mentioned, and the data obtained would be kept safe and used for only research purposes.

All the participants, including those who participated in the pre-pilot and pilot study, were given an information sheet which clearly explained the purpose of the study, and described the data collection procedures in detail (e.g. how and why this particular group of participants were selected and what was expected from them) (Appendix P). Formal consent was sought form each participant before starting the data collection (Appendix G).

6.8. Procedures

In this section, I explain the data collection procedures as to where and how the data were collected. This covers the context in which Study 2 took place, the participants' profile and why they were selected and the data collection procedures. Also, a detailed account of how the data was analysed, the steps followed, the measures selected for the analysis along with the issues and problems encountered in this process is presented.

6.8.1. Setting and participants

Researchers commonly select a method for participant sampling which is best suited to their research purposes. In this study, two sampling methods were applied to reach out the participants: convenience sampling and snowballing sampling. Convenience sampling method involves choosing the nearest individuals who 'happen to be available and accessible at the time' (Cohen et al., 2018, p. 218). In this method, researchers continue this process until they reach the required sample size. In snowballing sampling method, on the other hand, a small group of individuals who have the profile the researchers are interested in are identified and this group is used ' as informants to identify, or put the researchers in touch with others, who qualify for inclusion' (ibid.) Using these two sampling methods, the data in the current study were collected in the UK from L1 Turkish-speaking post-graduate students, many of whom had come to the UK on a scholarship program led by Turkish Ministry of Education. On this scholarship program, which is part of a larger project, the Ministry selects a number of candidates each year to study at different universities in Europe or the USA at post-graduate degrees. According to the candidates' selection criteria, all the bursaries must have successfully completed a bachelor degree program in Turkey, and should be under the age of 35. The bursaries are selected to study abroad at various disciplines on a contract with the Turkish Government, and the majority of these candidates go abroad for both masters and PhD degrees. However, some may return to their countries upon the completion of their master's degrees only.

Initially, 62 post-graduate students in the UK were recruited for the current study. At later stage, two of them were excluded since they did not participate in the whole data collection procedure. So, the data reported here comes from a total of 60 participants. While the majority of them (i.e. 55) were from aforementioned population of bursaries, some self-funded students or those studying under a different scholarship programme (e.g. those offered by their universities in the UK) were also included in the study. As mentioned earlier, prior to the data collection, the participants were given the background language questionnaire, which was modified for Study 2; this questionnaire was either sent through emails or handed to the participants in person, and allowed the researcher to obtain information about their educational background, amount of English language instruction they had received and the amount of time they had spent in the UK as well as their demographic profile (e.g. age, level of education, etc.). The advantage of obtaining such information beforehand was that the participants whose profile was not suitable for the study could be excluded from the study before the data was collected from them. For example, a couple of people stated that they were brought up as Kurdish-Turkish bilinguals and that their dominant language was Kurdish. These people were not included in the study to ensure the homogeneity of the group as much as possible.

As was stated in their responses to the questionnaire, the participants were aged between 22-35 (with a mean of 28) and their first language was Turkish. They all completed at least a bachelor's degree in their home country (i.e. Turkey), and all of them received formal English language instruction before coming to the UK. They reported that before coming to the UK, they had learnt English mainly through formal instruction and had about eight years of language instruction throughout their whole education life (starting from primary school, at the age of 9 or 10). However, in informal follow-up discussions, almost two third of the participants emphasized that English was taught as a foreign language in their country, and it was not widely, if not at all, spoken. In fact, the main use of English language was for communication purposes, i.e. with visitors or foreigners coming to the country. Therefore, the participants reported that the language education they had received throughout these years focused mainly on grammar instruction and/or it did not help them to develop L2 communication abilities.

The participants further noted that they had learned English mainly in intensive language courses either just before or after coming to the UK. As they mentioned, the Turkish Ministry of Education offered language courses to these candidates in Turkey before they made their applications to their degrees at universities abroad, so that they were able to obtain the required IELTs score (i.e. an international language test), which was one of the entry requirements of most universities abroad. If the candidates were not able to obtain this required score, they continued further to attend language courses after coming to the UK, commonly known as pre-sessional courses. Thus, although these participants had had long years of formal language education in their country before the start of their post-graduate degrees (i.e. master's degrees and/or PhD degrees), most of them learned English mainly through intensive language courses before and/or after coming to the UK.

The amount of time that the participants had spent in the UK varied between one month and five years, which was also another variable the current study was interested in. The participants also belonged to three levels of L2 proficiency, i.e. B1, B2 and C1 at CEFR levels, measured through the proficiency tests of OPT and EIT. One issue that is worth mentioning though, is that none of the participants used English for 100% of the time in their everyday life regardless of the time resided in the UK. All of them used English and Turkish almost equally in their lives in the UK; for example, at home, at the university, in social events or on social media. The reason for their frequent use of L1 was that they mostly had Turkish friends or were in constant interaction with a Turkish community, as they reported. In fact, it appeared that at several points in their life in the UK, they used Turkish even more than they did English.

6.8.2. Data collection

The data collection took place in two individual sessions on two separate days within the same week. As the piloting of the current study also revealed, the whole procedure took approximately 90 minutes per individual, which was considered a long period of time for the participants to sit and take the tests and the tasks. Similar to Study 1, in order to avoid participants' fatigue with a potentially damaging impact on their performance, it was decided that the data would be collected in two separate sessions.

Phase 1: The first individual meeting was arranged in a quiet office. In this first session, the participants were briefed about the aims of the study, why they were approached, their role in the study and what was expected from them in the whole process. They were given the project information sheet (explaining all such points and signed by the researcher) and a consent form for them to sign. Once their written consent was obtained, they were given the OPT. The researcher explained the test instructions (e.g. the format, all questions must be answered, etc.) in the participant's first language, i.e. Turkish. They were given 60 minutes to take the test. However, if they finished earlier, they were allowed to leave earlier.

In Study 2, for reasons of practicality and the availability of the participants, the participants had to take the test individually in the presence of the researcher, in contrast to Study 1, in which they sat the OPT in a classroom environment. Special care was taken in the current study, however, to make sure that the testing took place in an appropriate place, i.e. a quiet office, and that the testing settings were identical for each individual. Since most individuals finished the test in less than 60 minutes, the average time for this first session was 45 minutes. The second meeting was also arranged with each of them at the end of this session.

Phase 2: Within the same week, the individuals were met a second time in the same office. In this second meeting, they were first given two oral narrative tasks (Task A and Task B, Appendix M). The participants had been randomly divided into two equal groups: the first group was to narrate Task A in Turkish and Task B in English, while the second group performed the tasks in the opposite order. In other words, the tasks were counterbalanced. Table 6.3 below shows the design of the study, the tasks, and the counterbalance design.

Table 6.3. The groupings of the participants across tasks in Study 2.

	Order of the languages	
Tasks (N=60)	**Turkish-English**	**English-Turkish**
Task A (Best Friend) n=30	n=15	n=15
Task B (Quiet Please) n=30	n=15	n=15

As mentioned earlier, each oral narrative task contained 12 picture prompts. These pictures were colour-printed on an A5 paper. Although the task instructions were provided above the picture prompts together with the tasks and the characters' names (Appendix M) these were also explained orally to individuals.

After they completed the oral narrative tasks, the participants performed the EIT next. The procedures described in Gaillard and Tremblay (2016) (also explained in Study 1, Section 3.8.2) were followed in the EIT. Similar to Study 1, the test started with a practice session with five sentences in Turkish in order to make sure that the procedures were clear and understood by the participants. The test instructions were given in Turkish and included that 1) they were required to repeat each sentence as accurately as possible after they heard the beep sound, 2) they had only one attempt to listen and repeat the sentence, 3) they were not allowed to take any notes during the test. The test took each participants nine minutes to complete and their performances were audio-recorded.

Following the EIT, the participants were given a 10-minute break if they wished to have it. Then, they took the WMC tests starting with BWDS test. They were given BWDS first for practical reasons since the SNRT took a lot longer than BWDS and would potentially cause fatigue for some individuals. The English version of the BWDS was adopted from Awwad (2017) and included seven sets of increasing numbers; the first set started with three digits and the last set finished with nine digits. The researcher recorded each of these sets in English using the built-in 'Victoria, high quality' on a Macintosh computer. This feature within the computer read each digit at the rate of approximately one-digit 750ms. There was a delay between each set, and the length of the delay depended on the length of the set. Therefore, the length of the delays between the sets varied; for example, while there was a delay of approximately four seconds between the sets of three digits, there was a longer delay of approximately eight seconds between the sets of seven digits. As was discussed in Section 6.6, this test and its recording were tested in the pilot study to make sure that it worked well and was comprehensible.

The participants were told that they were going to be presented sets of increasing digits and were asked to listen to these sets and to repeat each set backwards. All the instructions were given in Turkish, i.e. their L1, and the test started with a short practice session. There were three sets of each digit group; i.e. three sets of four digits, three sets of seven digits, etc). If the participant failed to repeat successfully two of the three within each digit set, the test was terminated. The performances were audio-recorded; however, each participant's performance was also monitored by the researcher using the test sheet (Appendix N).

The BWDS was followed by the SNRT. The SNRT included lists of non-words, where the lists had varying lengths (i.e. eight lists of non-words at three list lengths; lengths of five, six and seven non-words). This test was adopted from O'Brien et al. (2006); however, the sets of the stimuli used in the test was originally taken from Gathercole et al. (2001). Each of the eight lists were made up of two sub-sets which were in either the same or different order. For example, a set of seven non-words, was followed by the same set of seven non-words; however, the second set had either the same or a different order than the first one. The order of the second sub-sets was randomly arranged (i.e. same or different order), except that the first and last pair of the items on both sets remained the same. With this arrangement, the participants were encouraged to process the whole stimulus and to reduce the eminence of the rearranged pairs of non-words. Also, the second presentations of the half of the eight lists were the same, i.e. it was identical to the first, and of the other half was different, i.e. the second presentation of the set had a different order of non-words. For example, while the first list was ' bordge, chud, nig, dack, keb, larm’, the second presentation of this set was ‘ bordge, chud, dack, nig, keb, larm’ (Appendix O).

Furthermore, all the sets were comprised of one-syllable non-words, each of which followed the ‘consonant-vowel-consonant pattern’, as already mentioned (O’Brien et al., 2006, p. 383). Within each list length (e.g. length of six non-words), all the vowel sounds were distinct of each other while the consonants were placed as phonologically different as possible. In the SNRT, the test takers were required to listen to these sets of non-words, i.e. words that have no meaning, and then to judge whether the second presentation of each was in the same order or not. The test also included a practice session, with four sets of four nonwords; two of these had the same order sub-sets and the other two had sub-sets of different orders. This practice session was important in that the whole procedure was well-understood by all the participants. The test lasted for seven minutes in total. The participants’ performance on the SNRT was audio-recorded and monitored by the researcher using the test sheet (Appendix O).

6.9. Data analysis

In this section, the data analysis procedures are described and discussed. These include the marking of the proficiency tests (i.e. OPT and EIT) and the WMC tests (i.e. SNRT and BWDS), the steps followed in the data transcription and coding and pause identification using PRAAT software (Boersma & Weenink, 2013).

6.9.1. Marking of proficiency tests

In the marking of the proficiency tests, the same procedures were followed as Study 1. The OPT was made up of 100 questions, and one point was awarded for each correct answer. Although, the participants were reminded to answer all the questions before the test started, in the cases of any unanswered questions, these were considered as incorrect answers, and therefore were not given any points. The maximum score one could obtain from the OPT was 100. Also, Cronbach alpha coefficient of .87 was obtained for the test, which indicated the test was highly reliable.

The EIT was also marked manually, listening to each participant's performance several times and marking each sentence they produced using the same scale which was described in detail in Section 3.9.1). In Study 1, the participants came from lower proficiency levels and learnt English in an EFL context. Therefore, some of them were not able to repeat all the sentences they listened to and wanted to terminate the test after a certain point, in which case the rest of the sentences were scored as zero. In contrast, in Study 2, which was carried out in a study-abroad context with participants of higher L2 proficiency levels, the participants continued with the test till the end, and therefore, they completed the test successfully.

For the scoring of the EIT performances, exactly the same procedures described in Study 1 was followed. These included obtaining the reliability statistics (i.e. Cronbach's alpha coefficient of .95) based on the participants' scores, converting the OPT scores into 120 for comparability reasons and creating a mean score of the OPT and EIT (50% of each). The participants' mean scores ranged between 44 and 103, and they were grouped into the CEFR levels based on these mean scores. A simple correlational analysis carried out between the OPT and the EIT scores showed that a high score on the OPT was associated with a high score of the EIT (r =.62, p < .001). The results also demonstrated that 38% of the variance in participants' OPT scores could be accounted for by their EIT scores. Similar to study 1, all calculations were done using relevant formulae on Excel and the groupings were based on the scoring system of Oxford Online Placement Test (Table 6.4). Readers could refer to Section 3.9.1 for a detailed explanation of the marking procedures of the proficiency tests and the groupings of the participants into different proficiency levels.

Table 6.4. The grouping of the participants across CEFR levels based on OOPT scoring system in Study 2.

CEFR Levels	Mean score of OPT and EIT*	n (N=60)
C2	81>100	-
C1	61-80	21
B2	41-60	21
B1	21-40	18
A2	1-20	-

**The maximum score is 120.*

6.9.2.Marking of working memory tests

Although the participants' performances were audio-recorded at the time of the WMC tests, their performances on both tests (BWDS and SNRT) were scored manually by the researcher during the tests. In each of the memory tests, the researcher monitored the participants' performances using the specific test sheet on each test (Appendix N and Appendix O, respectively) and for each individual. In the BWDS, the participants were given a score based on the last sets of digits that they were able to repeat successfully twice (Awwad, 2017; Awwad & Tavakoli, 2019; Wright, 2010). For example, if the person repeated successfully two of the three five-digit sets but could only repeat one of the three six-digit sets, then a score of 5 was awarded to this person. The total score one could obtain from this test was 9 (Table 6.5).

When it came to the scoring of the SNRT performances, the individuals were given a score based on the number of correct responses they provided. For example, a participant who gave 18 correct responses was awarded a score of 18. The maximum score that could be obtained from this test was 24 as there were eight lists of three lengths. In order to obtain a combined score for WMC, the scores from BWDS was converted into 24, then a mean score of BWDS and SNRT was calculated (50% of each). The combined score (out of 24) was used as a continuous variable for each participant. Further, a simple correlation coefficient between mean WMC scores and mean proficiency scores was calculated ($r = .32$, $p < .012$) in order to see how much language proficiency scores could be related to WMC scores, and the results demonstrated a weak correlation between them.

Table 6.5. Marking procedures of the WMC tests in Study 2

	Total score	**Marking procedure**
The BWDS (9 lists, each containing three sets of varying digit lengths)	9 (later converted into 24)	▪ Performances were monitored during the test and scored manually. ▪ Scores were given based on successful repetition of two (out of three) digit sets

The SNRT (24 sets of varying lengths of non-words, each containing two sub-sets)	24	▪ 1 point was given for each correct responses
Mean score	50% of each test	▪ A mean score was created for each individual

6.10. Measures adopted in the study

6.10.1. Fluency measures used in the study

Fluency measures adopted in Study 2 were limited to five measures in total; two breakdown measures, one repair, one speed and one composite measure. Study 1 aimed to explore L1-L2 fluency links adopting a wide range of fluency measures (a total of 17) to understand which measures were correlated between L1 and L2. However, Study 2 was extending the findings of Study 1 as far as the fluency aspect was concerned, and therefore, the range of fluency measures to be used was reduced and selected in a way that each aspect of fluency would be represented. For breakdown fluency, frequency of pauses, both at mid-clause and end-clause positions, were selected (Bosker et al., 2013; Huensch & Tracy-Ventura, 2017; Suzuki &Kormos, 2019). Fluency research has emphasized that an examination of pause location (i.e. mid-clause or end-clause) and pause character (i.e. filled or silent) is important for a better understanding of speakers' breakdown behaviour (Davies, 2003, De Jong et al., 2015; Hunter, 2017). Although some studies in this area considered only filled pauses (e.g. Lambert et al., 2017), some others combined both filled and silent pauses (e.g. De Jong et al., 2015; De Jong & Perfetti, 2011). Kormos (2006) argues that both could perform a similar function in speech in that they allow the speakers time to deal with processing demands such as planning content or language. However, research suggests that length and frequency of silent pauses (within AS-units) tend to change across different L2 proficiency levels. De Jong (2016), for example, reported that L2 proficiency had a mediating effect only for silent pauses within AS units but not filled pauses; more proficient learners produced fewer silent pauses. Therefore, I have followed this body of research and adopted silent pauses to represent breakdown fluency (i.e. number of mid-clause silent pauses and number of end-clause silent pauses).

As for the repair fluency, total number of repairs (i.e. repetitions, replacements, reformulations and false starts) was used since this measure has also been suggested as valid measure of repair behaviour (Kahng, 2014; Skehan, 2009) and the findings of Study 1 revealed a correlation for this measure. Articulation rate represented speed fluency as the only pure measure of speed. As was suggested by studies (e.g. De Jong, 2016; Hunter, 2017; Tavakoli et al., 2019, Suzuki & Kormos, 2019), this measure excludes pausing behaviour and therefore indicates how fast a speaker produces speech regardless of pausing phenomena. Finally, speech rate was included as a composite measure (combining breakdown and speed fluency) (De Jong, 2016; Huensch & Tracy Ventura, 2017; Mora & Frerrer, 2012; Tavakoli et al., 2019). A list of all measures used in the Study 2 is provided in Table 6.6 below.

6.10.2. Lexical complexity measures used in the study

In the present study lexical complexity has been used as covering different aspects of lexical knowledge (i.e. lexical diversity, lexical sophistication and lexical density) (please see Section 2.4.2 for a detailed discussion of each of these and the measures used to represent them). However, in the current study, I focused on lexical diversity (LD), which is ‘the range of different words used in a text, with a greater range indicating a higher diversity’ (McCarthy & Jarvis, 2010, p. 380), to represent lexical complexity 1) due to the limitations of the existing measures of lexical sophistication for analysis of Turkish language, i.e. although there exist some frequency lists in Turkish such as Turkish National Corpus (Aksan et al., 2012), most of these are based on written data and are quite limited in their word range (less than 50 million), and 2) due to the time constraints for an analysis of lexical density (i.e. the ratio of content words to function words) for both languages. Therefore, given these limitations, an analysis of both languages in the current study was only possible by using LD measures.

TTR was once the most widely used measure of LD (i.e. traditional method) in the field and across disciplines. However, as discussed in Section 2.4.2.2, TTR was also known for its being sensitive to the text length (i.e. as the text lengths becomes longer, TTR value decreases). Measures proposed later (e.g. D measure, HD-MLTD or Guiraud), which are also essentially based on TTR values, have been developed to overcome the issue of text-length sensitivity through different random samplings of TTR curves. Although the text-length sensitivity still remains an unsolved problem in this area (Treffers-Daller, 2013; Treffers-Daller et al., 2018), these measures have been employed in a number of studies to measure the complexity in texts (Daller & Xue, 2007; Tidball & Treffers-Daller, 2008). Some researchers in the field of CALF also used LD measures to build up a better picture of the complexity of speech production or texts (e.g. Awwad, 2017; Revesz et al., 2016; Suzuki &Kormos, 2019).

Further, recent studies argue that although there has been tremendous efforts in developing new measures, simple counts of number of different words (i.e. types) in a text can sometimes yield more accurate and reliable results than complex formulae such as those in the aforementioned measures (e.g., Treffers-Daller et al., 2018), provided that the text length (i.e. the number of words analysed) is kept constant across all texts (Lancashire & Hirst, 2009). In the absence of one single valid measure, Treffers-Daller et al. (2018) suggests that lexical should use both a traditional method, i.e. TTR values, and later methods. Following from these suggestions, two measures were selected to be included in the current study: D measure (as a more recent method) and simple TTR values (as the traditional method) (Table 6.6). In order to obtain TTR values VocabProfile (**www.lextutor.com**) software was used while D values were computed through D tool in Lognostics ((**www.lognostics.co.uk**).

D measure, essentially based on a formula that models falling TTR curves with increasing text length, is calculated through a random sampling of different text lengths. In the current study, the D tool in Lognostics software was used since the program allowed for analysis of small texts, i.e. texts longer than 50 words, and for examining the LD in different languages (i.e. Turkish and English in this case). Although I fully acknowledge that D measure has not gone uncriticized like all other measures, it was adopted in the current study for two reasons 1) a number of previous studies demonstrated the use of this measure (e.g. Malvern et al., 2004; Revesz et al., 2016; Yu 2010), and 2) given that the speech samples in the current study ranged between 70 and 150 words, D was a more suitable measure to examine the LD than other measures such as MLTD, which required a minimum of 100 words (McCarthy & Jarvis, 2010).

Table 6.6. Measures of fluency and lexical complexity used in Study 2.

Aspect of oral performance	**Measures**	**Definitions**
Fluency (Kormos, 2006)	**Breakdown fluency**	
	No of mid-clause silent pauses (MCSP)	The total number of silent pauses within the clauses divided by the total speaking time
	No of end-clause silent pauses (ECSP)	The total number of silent pauses between the clauses divided by the total speaking time
	Repair fluency	
	Total number of repairs	The total number of repetitions, replacements, reformulations and false starts
	Speed fluency	
	Articulation rate	Number of syllables produced divided by total speaking time (excluding pause time) in seconds
	Composite measures	
	Speech rate	Number of syllables produced divided by total speaking time (including pause time) in seconds
Lexical complexity (represented through LD)	D Measure	Adjusted type-token Ratio (TTR) computed through Lognostics tool (Meara & Miralpix, 2007).
	TTR	Simple calculation of type-token ratio (TTR) which represents the ratio of different words used to the total number of words.

6.11. Data transcription and coding

In this section, I describe the ways in which the data collected through oral narrative tasks in both L1 and L2 were analysed. I should point out that the same procedures described in detail in Study 1 (Section 3.10) were followed in the data analysis of Study 2. These include all the steps of the data analysis for fluency performance; i.e. the data transcription, coding and annotation for fluency measures. An addition to these procedures in Study 2 was the analysis of the data for lexical measures. As such, in this section, while readers are referred to the relevant section in Study 1 for detailed accounts of the data for fluency measures with only a brief account of these provided, a detailed description of the data analysis for lexical performance is presented.

6.11.1. Simple transcriptions

Similar to Study 1, speech samples elicited from the participants through audio-recordings of their performance in the narrative tasks were transcribed using Transcriber (2005), an online software. However, since the participants in study produced longer speech samples than those in Study1, only the first 60 seconds of their speech were used for analysis. For each participant, a separate word document, including the simple transcriptions of L1 and L2 speech samples was created. The transcriptions started with the first utterance the participants produced, excluding any silences at the very beginning and included every utterance produced with no punctuation marks (please see Section 3.10.1 for a detailed account of the simple transcriptions). While filled pauses were noted down in the simple transcriptions, silent pauses were not identified at this stage. Once the transcriptions were completed, the word files were copied into a different folder and these new copies were used for further analysis, which would also be helpful to do the syllable counts and lexical analysis at a later stage.

6.11.2. Coding for units of analysis: AS-units

Following the transcriptions, the data were coded into AS-units. As was in Study 1, AS-units were used to segment the data, following from Foster et al., (2000). In Section 3.10.2, different unit types were already discussed, and the definition of the AS-unit was provided. Therefore, it would be sufficient to note here that an AS-unit consists of one main clause and one or more subordinate clauses, if any. Elliptical expressions, such as ‘yes,’ really’ ‘perhaps’ are also counted as AS-units, or part of an AS-unit when these start an utterance. These units have been commonly used in recent studies of this kind (e.g. Awwad & Tavakoli, 2019; De Jong et al., 2015; Hunter, 2017; Malicka & Levkina, 2012; Tavakoli & Foster, 2011) as a valid and reliable unit of analysis for oral data, and therefore AS-units were considered as the unit of analysis in the current study as well.

The coding process required listening to the recordings a few times to ensure that the units were in line with the intonation patterns of the speakers. This was particularly essential for the analysis of Turkish speech samples in the current study because in Turkish, the subject position in a sentence could change; it could come at the end of the sentence or in the beginning. It is also possible that the subject could be omitted if the verb is already indicative of the subject. Please see the subject position (i.e. kedi, which means 'cat') in the sentences below, for an example of this:

1. | Daha sonra tüfeği alıyor *kedi* | yine fareyi yakalamaya çalışıyor |
   | Then *the cat* takes the riffle | (*the cat*) tries to catch the mouse again |
2. | Daha sonra tüfeği alıyor | *kedi* yine yakalamaya çalışıyor |
   | Then (*the cat*) takes the riffle | *the cat* tries to catch the mouse again |

In the first example (1), the subject belongs to the first clause, and it is omitted in the second clause. Whereas in the second example (2), subject omission is observed in the first clause, and the subject, i.e. *kedi,* belongs to the second clause. As these examples also illustrate, the identification of subject positions was only possible through multiple listening to the recordings, and without a consideration of the intonation patterns in the speech, it would be impossible to identify the clause boundaries.

6.11.3. Classification of repair fluency

The same procedures in Study 1 were followed in the classification of repair instances. After procedures for the coding of AS-units, repair instances (repetitions, replacements, reformulations and false starts) were identified and marked with a set of conventional symbols (Appendix H). The instances of each repair type were counted for each speech sample (in both languages), and then using the relevant formula in the excel file, these total numbers of instances were averaged to 60 seconds when the total speech length in a given sample was less than 60 seconds. The definition of each repair type can be found in (Table 3.6) and was also explained in Study 1 (Section 3.9.2). Inter-rater reliability for the classification and the placements of the repair types is discussed below in Section 6.12.

6.11.4. Identification of pauses: using PRAAT

While the filled pauses were already identified in the simple transcription stage, silent pauses and lengths of all pauses were identified using PRAAT software (Boersma & Weenink, 2013). PRAAT is a tool that serves an objective way for researchers to make an automatic analysis of certain measures such as number or lengths of silent pauses. It has been used widely in a number of fluency research (e.g. Cucchiarini et al. 2000, 2002; De Jong & Perfetti, 2011, De Jong & Wempe, 2009; Hunter, 2017) and was used in Study 1 as well. Therefore, the same procedures used in Study1 for PRAAT analysis were followed in the current study (please see Section 3.10.4 and Table 3.7 for a detailed explanation of the procedures and further information about PRAAT). These could be summarized in a few steps: 1) audio files were converted to .WAV sound format, 2) using the Annote to Texgrid (silences) command, text grids with segments and tiers for filled and silent pauses were created for each audio file, 3) silences of .25 of a second or longer (e.g. De Jong & Bosker, 2013; De Jong et al., 2015, Huensch & Tracy-Ventura, 2017) were marked against the spectrograms after multiple listenings and checks for their location and precise lengths and 4) the total numbers of all pauses, together with their lengths, were noted down on each participants' word file, which would be help at a later step to make further analysis of mid-clause and end-clause pauses. A total of 120 speech samples were analysed in this way, i.e. 60 participants producing speech in both Turkish and English, and 10% of the data was re-analysed for intra-rater reliability (please see Section 6.12 below).

Please note that unlike in Study 1, a couple of issues emerged in this process (i.e. using PRAAT to identify pause lengths and numbers) in the current study. These are worth mentioning here as they will further contribute to our understanding of the whole process and make the analysis clearer for future studies. One issue was that when a filled pause was immediately followed by a word and the last sound in the filled pause almost combined with the beginning sound of the following word (e.g. *erm made it*), it was rather difficult to identify where exactly the pause ended and the word started. In Figure 6.1**.** below, the highlighted segment (i.e. indicated as sounding), demonstrates such an instance, i.e. a filled pause and a word combined. Although such cases were rare, when observed, the ending of the pause and the start of the word had to be marked intuitively again through multiple checks and listening to the sound files.

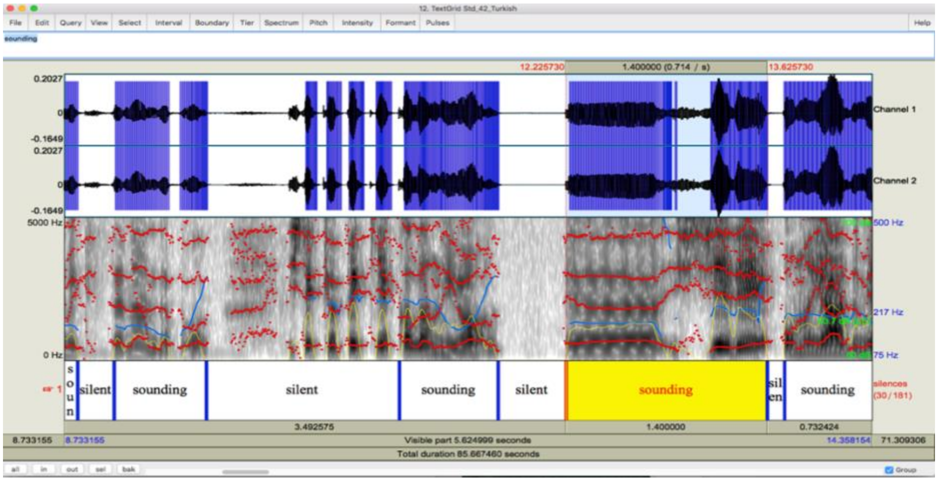


Figure 6.1. PRAAT sound file and text grid illustrating a filled pause followed by a word

A thorough examination was also needed when words ended in plosives (i.e. voiced p, t or k or voiceless b, d, or g), especially when these were proceeding vowels. A plosive is a consonant sound. When it is produced, the flow of air is blocked with lips pressed together, or teeth or palate. When the air is released, the sound is produced in an explosion (Roach, 2009). It was observed that when the speaker uttered words ending in these sounds (e.g. *cat*, *dog* in English or *yapip* (doing), *carpip* (hitting) in Turkish), these words were indicated by PRAAT within the sounding segment, as should be expected. A closer inspection, however, revealed in most of the occasions that the word actually ended at a longer distance than the end of the sounding segment, and this resulted in a different pause location and length. Again, the identification of the end of such words was not an easy task. The approach taken here was to listen to the sound file several times and mark the ending at the most appropriate place on the spectrogram, making sure that the word is fully pronounced.

In Figure 6.2 below, an example of this can be seen. Although the highlighted segment is indicated as sounding (the word produced is ‘*dog*’), the words actually ended at the end of the next segment, which is indicated as silent. These issues further indicated that much as PRAAT served as a very useful and objective tool for the identification of pauses, a careful manual examination (which is, by all means, a very time-consuming process) was required through listening to the sound files and zoomed-in versions of the text grid.

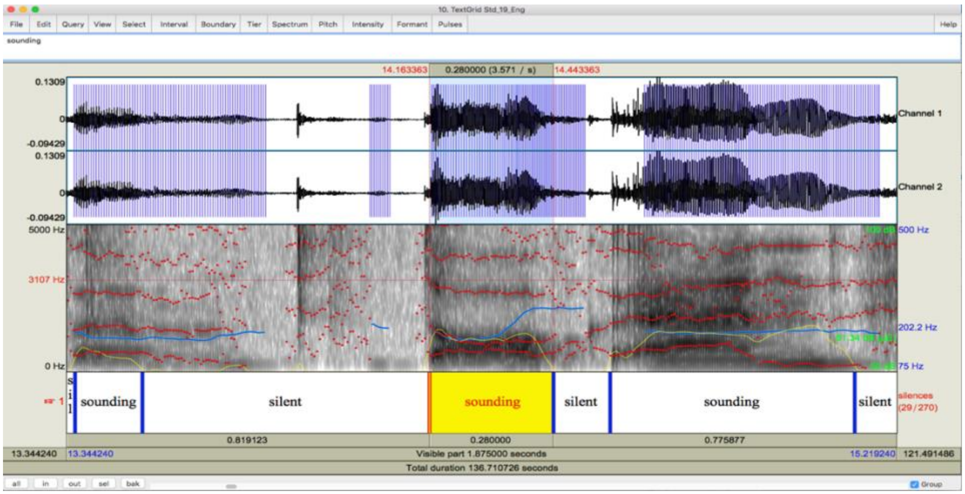


Figure 6.2. PRAAT sound file and text grid illustrating a word ending in a plosive (i.e. t)

6.11.5. Syllable counts

As was the case in Study 1, syllable counts, rather than words counts, were used to quantify the speech in Study 2. Syllable counts were required in this study as well, because 1) words counts may not be produce reliable results when comparing two structurally different languages (i.e. Turkish and English), 2) in order to be able to compare the results with those from other similar ones, which also adopted syllable counts (e.g. De Jong et al., 2015; Huensch & Tracy-Ventura, 2016, 2017), syllables counts were needed and 3) the calculation of some fluency measures adopted here (e.g. articulation rate or speech rate) was based on syllable counts.

In Study 1, the syllable counts in English were done using a syllable counter website (**www.syllablecount.com**), following from Thai and Boers (2016), whereas for those in Turkish samples, manual counts were used. In Study 2, while the calculations were done using the same tool in English, Turkish counts as well were carried out using a similar programme, Hesapla Online (**www.hesapla.online.com**). To ensure the reliability of both tools for each language, 10% of the calculations carried out by each tool were double checked by manual counts, and an inter-rater reliability of above .98 (Cohen's Kappa) was achieved (also see Section 6.12).

In order to count the syllables, the simple transcriptions were re-visited, and these were copied into a different folder to clear the data from any filled pauses. Once this was done, each transcript was copied and pasted into the relevant column in the counter (i.e. Syllable Counter for English and Hesapla Online for Turkish). After ensuring that all words were identified or found in the dictionary of the tool, the total number of the syllables was noted down once again on the coded sheet of each participant. It should also be noted that unpruned data (i.e. the data including the repair phenomena) was used for syllable counts.

6.11.6. Lexical complexity (LD) analysis

LD analysis in the current study was based on the use of two measures; namely a traditional method, *simple type-token ratio,* and a relatively more recent measure, *D measure*. As discussed in Section 5.6 earlier, several issues should be kept in mind when carrying out any lexical analysis. To start with, the operationalization of the basic unit of analysis, i.e. what it is counted as a type (i.e. a different word) and what is not, should be defined very clearly. Different approaches have been adopted by researchers so far regarding the operationalization of types; while some researchers considered all inflected forms of a word (e.g. buy, buying, buys) as tokens of the same type (Jarvis, 2002), some others took these as different types (e.g. Yu, 2010). However, the research in this area further suggest that such different approaches are likely to result in different LD scores and could potentially overestimate or underestimate the scores (Treffers-Daller & Korybski, 2015). Therefore, the approach adopted in any lexical study and the decisions taken, if any, should be clearly defined at the outset so that the findings could be interpreted with a consideration of these.

Another issue concerns whether the data used is lemmatized or not before the data is subjected to any analysis. Treffers-Daller et al. (2018) caution that this is a crucial factor in lexical analysis, particularly, in cross-linguistic studies such as the current study, where languages under investigation are typologically different. Further to the issue, sensitivity to the text length still remains as a limitation in lexical studies, regardless of which recent measure was used, i.e. those which have been proposed to reduce the effect of text length (e.g. D measure, HD MLTD or Guiraud) (Section 5.6). Therefore, it is recommended that the text length should be kept constant across the texts (Lancashire & Hirst, 2009; Treffers-Daller et al., 2018).

Taking these issues into consideration, the current study aimed to analyse LD (which represents lexical complexity) by keeping the text length constant, at 100 words per text. This number was selected because most of the speech samples in this study were around this word range. However, a limitation that I should acknowledge here was that a couple of the samples (i.e. six English samples and three Turkish samples) fell below this range (i.e. ranging between 70- 80 words), therefore these samples had to be approximated to 100 words after D and TTR values were obtained for a standardization of all the scores.

Before the lexical analysis was carried out, the texts were prepared for analysis. The simple transcriptions were re-visited at a third time and these were re-created under a different folder. According to the recommendations in Treffers-Daller et al. (2018), the first step in this process should be the data clearing; therefore, each transcript was cleared from any proper names (e.g. tom, jerry), filled pauses and repair phenomena (i.e. repetitions, false starts). With regard to the repetitions, only doubling words in Turkish, one of the important features in Turkish, (e.g. *adim adim* (step by step) or *ağır ağır* (slow and gradual) were not removed from the data because the reason for using the same word again in such cases was obviously not related with repair behaviour, but rather these were reduplicated to intensify the meaning.

As a next step of the data preparation, lemmatization was carried out. Again following from the recommendations of previous research, the basic unit of analysis (i.e. type) in the current study was defined; all inflected forms of one word (e.g. tries, trying, tried or working, works, worked) was considered as one type whereas derivational forms were taken as different words (e.g. worker or workable) (Jarvis, 2002). Since the languages under investigation (i.e. English and Turkish) are two typologically different languages, a few words must be said here about analysis of Turkish. Turkish is an agglutinative language and thus, the word structures are formed by adding morphemes (suffixes) to root words. Some of the inflectional suffixes include plural forms (-lar/ler), possessive forms or relative forms whereas derivational suffixes include those which change the word family or make a new word (e.g. those turning a noun into another noun or those turning a noun into an adjective). Accordingly, all Turkish words formed by derivational suffixes were taken as different words, as was the case in English.

The lemmatization was carried out manually in this study. At this point, one could argue that the lemmatization should be done using a software such as Natural Language Tool Kit (NLTK) (Bird, Klein, & Loper, 2009) or the appropriate command in CLAN software (MacWhinney, 2000), both of which could be considered as reliable tools in data lemmatization for English language. However, for reasons of comparability between the two languages as well as time constraints, manual lemmatization was deemed more suitable. In addition, the texts to be lemmatized were relatively short (ranging between 70- 100 words), so manual lemmatization did not a pose a problem.

After the lemmatization and the data clearing process, the first 100 words were copied and pasted into these programs; first into the D tool in Lognostics (Meara & Miralpeix, 2007, 2016), then into VocabProfile (Cobb, 2008). The scores generated by the tools were noted down into an excel file for further statistical analysis.

6.11.7. Combining all scores for statistical analysis

The final step of the data analysis was compiling all the scores and the measures onto one sheet for statistical analysis. Part of this was already done during the process; the coded sheets created for each individual contained information about the number of AS-units and clauses, the number of all repair types (i.e. repetitions, replacements, reformulations and false starts) and the precise lengths of speech samples in both L1 and L2 (also can be seen in Table 6.7). Once all pauses were identified in PRAAT and transferred into these coded sheets, the number of mid-clause and end clause pauses were calculated, and these too were noted down on the same sheets.

All of these measures were transferred into an excel file. Total number of syllables, the scores from the proficiency tests and the working memory tests, and the lexical complexity scores had already been noted on this excel file. Therefore, the file was ready for the application of some formulae regarding some of the fluency measures. For example, for the calculation of speech rate, the formula below was used:

$$\frac{number\ of\ syllables\ \ x\ 60}{total\ speech\ time\ (secs)}$$

After all measures were calculated, the scores on the excel file was transferred into SPSS for further statistical analysis. Table 6.7 below presents a summary of all procedures regarding the data transcription and coding and data preparation for statistical analysis.

Table 6.7. A summary of the data transcription and coding procedures in Study 2.

Stages of data preparation for statistical analysis	Steps followed
	Fluency analysis
Simple transcriptions	▪ **A word file** (**copy 1)** was created for each participant including L1 and L2 speech ▪ Speech samples were transcribed (Transciber (2005)) with filled pauses and no punctuation marks (*identification of filled pauses*) ▪ Simple transcription files (copy 1) copied into another folder **(copy 2)**
Coding for AS-units	▪ The data were segmented into AS-units on new copies **(copy 2/coded sheets)**
Classification of repair fluency	▪ Repetitions, replacements, reformulations and false starts were marked with a set of conventional symbols on the coded sheets (copy 2). (*identification of instances of repair fluency)*
Identification of pauses: using PRAAT	▪ 120 audio files (Turkish and English, 60 each) were converted into .WAV format ▪ A text grid was created for each audio file (Annotate to Textgrid (silences) command) ▪ Tiers were added against the spectrogram: filled and silent pauses ▪ 60 seconds of each audio file was analysed: - Beginnings and endings of silent and filled pauses were marked. (*identification silent pauses and lengths of all pauses*) ▪ Coded sheets (copy 2) were re-visited to complete or correct any missing information ▪ Number and lengths of pauses were noted down on the coded sheet (copy 2).

	▪ Number and lengths of mid-clause and end-clause pauses (filled and silent) were calculated and noted down on the coded sheet (copy 2) (*identification of mid-clause and end-clause filled and silent pauses*)
Syllable counts	▪ Simple transcriptions were re-visited and copied into another folder (**copy 3)** ▪ New copies were cleared from any filled pauses ▪ Syllable counts were performed: - Syllablecount.com was used for English speech samples - Hesaplah.online.com was used for Turkish speech samples ▪ Number of syllables were noted down on the coded sheets (copy 2).
Lexical complexity analysis	
Data preparation	▪ Simple transcriptions (copy 1) were re-visited and new copies were created under a different folder **(copy 4).** ▪ Data were cleared from proper names, filled pauses and repairs (i.e. repetitions, replacements, reformulations and false starts) (copy 4) ▪ Data were lemmatized - All inflected forms of one word were marked as different words, and derivational forms were marked as the same word.
Performing LD analysis	▪ The data (copy 4) were analysed for D measure and TTR values - Data were copied and pasted into Lognostics (D measure) - Data were copied and pasted into Vocabprofiler (TTR valuees) ▪ The lexical scores were noted down onto an excel file (Ex1).
Combining all scores for statistical analysis	

Transferring all the scores and measures onto the excel file and SPSS	▪ Coded sheets (copy 2) were re-visited and fluency measures were transferred into the excel file (Ex1). - Numbers and lengths of filled and silent pauses (mid-clause and end clause). - Number of repairs - Number of syllables - Total speech time ▪ Remaining fluency measures (e.g. articulation rate, speech rate) were calculated using the relevant formulae in Excel. *(identification/calculation of speech and composite fluency measures)* ▪ Fluency measures were approximated to 60 seconds, and lexical measures were approximated to 100 words, where applicable. ▪ Scores from proficiency tests and WMC tests were transferred into the excel file (Ex1). ▪ All scores were transferred into SPSS.

6.12. Inter-rater/inter-coder reliability

The reliability of the coding and the analysis of the data was checked in a similar way to Study 1; i.e. a sub-set of the data (ranging between 10% to 20%of the data) was re-analysed at different phases of the procedure. Help was sought from the same experts who did the re-analysing in Study 1 since they had already received necessary training and therefore were assumed to be familiar with the marking procedure. To start with, 10% of the OPT sheets and 20% of the EIT performances were re-marked by the same second expert; with Cohen's Kappa of 1.00 and a simple correlation coefficient of .96 being achieved, for OPT and EIT performances, respectively, between the sets of scores given by the two markers.

Next, 20% of the coded sheets were checked by a second expert for the identification and the placement of AS-units, the clause boundaries and the repair instances (i.e. repetitions, replacements, reformulations and false starts). When Cohen's Kappa statistics were carried out to determine the consistency between the coders, a reliability of above .95 was achieved for each of these (e.g. identification of AS-unit boundaries) (see Table 6.8 for each of these statistics). Further, the reliability for syllable counts, which were done using online tools, was checked by manual counts of the researcher in 10% of the samples. The agreement between counts of the researcher and the tools was 100% (Cohen's Kappa of 1.00).

As for the identification of silent pauses and the pause lengths using PRAAT, the researcher analysed 10% of the data a second time. Pearson product-moment correlation coefficient between the first and second coding was calculated on the placement of silent pauses and the identification of pause lengths, and a correlation of above .95 for each of these was achieved.

Finally, 10% of simple transcriptions (after the data clearing process) was checked by a second bilingual after a training session on the identification of types and tokens was provided. Combining all the data, Cohen's Kappa was performed on the number of different words in both languages. An inter-coder reliability of above .92 was achieved.

Table 6.8 presents all the inter-rater/coder reliability calculations across different phases of the data analysis.

Table 6.8. Inter-rater/inter-coder reliability across different phases of the data analysis in Study 2

Segments of the data re-analysed/coded (in both languages where applicable)	**Second-rater/coder and reliability**
The scoring of the OPT (10%)	▪ An expert in the field ▪ Inter-rater reliability: ▪ Cohen's Kappa of 1.00 (100%) combining the data
The scoring of the EIT performances (20%)	▪ A trained second rater ▪ Inter-rater reliability: ▪ Pearson correlation coefficient of .98 between the sets of scores
Coding of the transcriptions (20%) on *the identification* and *the placement* of: • AS-units • clause boundaries • filled pauses repairs (i.e. repetitions, replacements, reformulations and false starts)	▪ A second bilingual expert in the field. ▪ Inter-coder reliability: - Combining the data, Cohen's Kappa for the identification of: ○ AS-units and clause boundaries (1.00) ○ filled pauses (1.00) ○ repairs (.95) - Combining all data, Cohen's Kappa for the placement of: ○ AS-units and clause boundaries (.95) ○ filled pauses (1.00) ○ repairs (1.00)
Syllable counts by the online tools (10%)	▪ The researcher (manual checks) ▪ Inter-rater reliability: ▪ Cohen's Kappa of 1.00 (100% agreement)
PRAAT analysis on (10%): - The identification and the placement of the silent pauses The identification of the lengths of all pauses (i.e. silent and filled pauses)	▪ The researcher (analysis of the data a second time) ▪ Intra-rater reliability: - Pearson correlation coefficient between the first and second analysis for the *identification* of:

	▪ silent pauses (1.00) ▪ lengths of silent pauses (0.95) ▪ lengths of filled pauses (0.96) ▪ Pearson correlation coefficient of 1.00 between the first and second analysis for the *placement* of silent pauses
Lexical Analysis (10%): The identification of the types/ different words	▪ A trained second bilingual expert in the field ▪ Inter-coder reliability (on the number of types): ▪ Cohen's Kappa of above 0.92 (combining all the data)

6.13. Summary of the chapter

In this chapter, I have outlined the research design and the procedures followed in the data collection and the data analysis. I began with an explanation of the research aims and presented the research questions and the hypotheses, which guided Study 2. Then, I explained the design a detailed explanation of the pre-pilot and pilot studies carried out before the start of Study 2. I set out the details about the particular context in which Study 2 was carried out, the participants of the study and the steps of the data collection, along with a rationale for the choices made. Finally, I described in detail how the data were analysed and prepared for statistical analysis. In the next chapter, I present the results of the statistical analysis carried out for Study 2.

CHAPTER 7. STATISTICAL ANALYSIS AND RESULTS: STUDY 2

7.1. Introduction

This chapter provides the details of statistical analyses run in order to answer the RQs of Study 2. The results of the analyses are reported in relation to each RQ and are presented in a systematic and detailed way under two main sections. The first section deals with the statistical analysis of fluency measures and the independent variables of Study 2 while the second section deals with the statistical analysis of LD measures. In each section, the preliminary analysis of the data is also provided including the descriptive statistics and the visuals for the assumptions for the inferential statistics.

As described in the previous chapter, Study 2 had a mixed between-within factorial design to investigate whether there is a relationship between L1 and L2 oral performance, regarding fluency and lexical complexity aspects, and whether the relationships between L1 and L2 fluency behaviour are mediated by individual factors of L2 proficiency, WMC and LoR. L1 and L2 oral performances (represented through fluency and LD measures) were dependent variables of Study 2 whereas the independent variables were L2 proficiency (described with three levels), WMC and LoR as continuous variables. Five fluency measures and two LD measures were used in both L1 Turkish and L2 English, also as continuous variables, to operationalize fluency and lexical complexity behaviours, respectively.

7.2. Data screening and cleaning for errors and/or outliers

In order to answer the RQs of Study 2, a number of statistical analysis were run using IBM SPSS 23.0. Similar to the procedures adopted in Study 1, data files, which were initially created in Microsoft Excel Program and contained all the fluency and LD measures, were transported into SPSS for inferential analysis. Before any statistical analyses were carried out, the data were screened for any errors which may have occurred in the data entry (Pallant, 2013). A couple of errors were detected at this stage, which included some missing values in the excel file for *total repair* in both languages and *number of mid-clause silent pauses* in Turkish. Therefore, the coded speech samples were checked back again for these missing values. The data in the excel file were also checked for the specific formulae for such measures as speech rate or articulation rate since an error in any of these would be crucial, resulting in a totally different value.

As another step of the data screening, it was ensured that there were no outliers, i.e. values which are way below or way above than the other scores. The outliers for both dependent and independent variables were checked by using the 'Explore' procedure in 'Descriptives' in SPSS. At this stage, one extreme outlier was detected for the fluency measures of speech rate and articulation rate in both languages. Before taking a further decision about the outlier's data, it was checked whether the figures for these measures were genuine or not (i.e. whether an error or not), by checking back with the participant's data. A re-inspection of the data indicated that there was another error in the data entry for the syllable counts, which had indisputably affected the figures on both speech rate and articulation rate. Once these numbers were corrected, the data set was checked for outliers once more time by using the same SPSS command, and none was observed.

7.3. Analysis of fluency measures

7.3.1. Preliminary analysis for fluency measures

Once it was ensured that the data file had no errors or outliers, the next phase was to run the descriptives statistics, which were needed to describe the variance in the performance of the sample and to draw a broad picture about the group differences . Given that all measures were continuous variables in the study, the Descriptives command in SPSS was used. This produced summary statistics of mean, median and standard deviation, for all measures. Further to the descriptive statistics, paired-samples t-tests were also carried out in order to see whether the group differences were statistically significant, and the effect sizes (Cohen, 1988) were calculated to assess the magnitude of the effect. I should note that similar to Study 1, the effect sizes were interpreted using the field-specific criteria in Plonsky and Oswald (2014) who proposed benchmarks of d = .60 (small), d = 1.00 (medium) and d = 1.40 (large) for within group comparisons.

Table 7.1 presents the results obtained from the descriptive statistics and paired-samples t-tests for the fluency measures in Turkish and English (please note that '*df*' value is always 59). As can be seen in the table, the figures for fluency measures in both languages were different for all measures of breakdown, repair, speed and composite measures. There was an increase in the mean scores from Turkish to English for *number of mid-clause silent pauses* (M = 11.72, SD = 4.17 and M = 22.34, SD = 5.37) and *total repair* (M = 3.31, SD = 2.42 and M = 7.00, SD = 4.25). The observed increase from Turkish to English for both measures were statistically significant, with t = -14.87, p < .001for *number of mid-clause silent pauses* and t = -7.40, p < .000 for *total repair*. Cohen's statistics indicated medium to large effect sizes for these measures (d = 2.20 and d = 1.06, respectively) (Plonsky & Oswald, 2014).

For other fluency measures, the figures indicated a statistically significant decrease, with large effect sizes, from Turkish to English for *number of end-clause silent pauses* (M = 13.26, SD = 3.25 and M = 9.11, SD = 2.81, with t = 7.93, p < .001, d = 1.36), *articulation rate* (M = 386.11, SD = 39.34 and M = 213.86, SD = 31.40, with t = 37.41, p < .001, d = 4.83) and *speech rate* (M = 249.24, SD = 39.24 and M = 106.39, SD = 28.21, with t = 28.86, p = .001, d = 4.18). The large effect sizes for these measures further suggest that the differences between fluency measures in L1 and L2 are important. With such low p values, type 1 error (i.e. the rejection of a true null hypothesis) is very unlikely.

The means and standard deviations obtained from the descriptive statistics provide a general overview of the data. The lower figures for *number of mid-clause silent pauses* and *total repair* and the higher figures for *speech rate* and *articulation rate* in L1 suggest that speakers were overall more fluent in their L1. The observed higher mean for *number of end-clause silent pauses* in Turkish indicated that the participants produced a higher number of end-clause pauses in their L1. This figure, i.e. higher Turkish number of end-clause pauses, could also be suggested to indicate more fluency in L1 since it could be related to the distinction between native speakers' and L2 learners' pausing behaviour within and between clauses. To put it more clearly, it is suggested that L2 learners pause more within the clauses whereas L1 speakers tend to pause more often at clause boundaries, and *higher number of end-clause silent pauses* in L1 in this case seems to support this suggestion.

Table 7.1. Descriptive statistics and t-tests for L1 and L2 fluency measures in Study 2

Fluency measures	Minimum	Maximum	Mean	SD	t	Sig. (2-tailed)	CI (95% for Mean)		Cohen's d
Breakdown measures									
Tur number of MCSP	3.00	20.00	11.72	4.17	-14.87	.001	Lower Bound	10.64	2.20
							Upper Bound	12.80	
Eng number of MCSP	11.00	36.00	22.34	5.37			Lower Bound	20.95	
							Upper Bound	23.73	
Tur number of ECSP	5.31	20.00	13.26	3.25	7.93	.001	Lower Bound	12.42	1.36
							Upper Bound	14.10	
Eng number of ECSP	4.00	17.00	9.11	2.81			Lower Bound	8.38	
							Upper Bound	9.84	
Repair measures									
Tur total repair	.00	11.00	3.31	2.42	-7.40	.001	Lower Bound	2.68	1.06
							Upper Bound	3.93	
							Lower	5.90	

Eng total repair	.00	19.00	7.00	4.25			Bound		
							Upper Bound	8.10	
Speed measures									
Tur articulation rate	324.27	504.70	386.11.	39.34	37.41	.001	Lower Bound	375.95	4.83
							Upper Bound	396.28.	
Eng articulation rate	130.62	292.48	213.86	31.40			Lower Bound	205.75	
							Upper Bound	221.97	
Composite measures									
Tur speech rate	166.00	333.67	249.24	39.24	28.86	.001	Lower Bound	239.10	4.18
							Upper Bound	259.38	
Eng speech rate	45.00	179.00	106.39	28.21			Lower Bound	99.10	
							Upper Bound	113.68	

7.3.2.Preliminary analysis for independent variables

Descriptive statistics were also run for other variables of the study; i.e. L2 proficiency scores (EIT scores, OPT scores and a mean score of the two), WMC scores (BWDS, SNRT and a mean score of the two) and LoR. Table 7.2 below presents an overview of the descriptive statistics for these measures. The EIT scores ranged between 30 and 103, with *M* = 70.35, *SD* =19.54, the OPT scores between 50 and 103, with *M* = 75.13, *SD* = 12.94, and the mean proficiency scores between 44 and 103, with *M* =72.87, *SD* = 14.74. These figures indicate that although the variance of the scores was smaller in the OPT, overall, the participants performed similarly in both language tests.

When we turn to the descriptive statistics for WMC scores, the figures indicated that overall, the participants performed better in SNRT (*M* = 15.62, *SD* = 2.93) than in BWDS (*M* = 4.25, *SD* = .96). The scores ranged between 3 and 8 for BWDS and between 10 and 22 for SNRT. The mean score for WMC was 8.31 and individual scores for the mean of WMC ranged between 5.36 and 11.72. Finally, the average amount of time the participants spent in the UK was 18.67 months, indicating that majority of them had more than one year of experience of living in the UK (Table 7.2).

Table 7.2. Descriptive statistics for independent variables in Study 2

Independent variables	Minimum	Maximum	Mean	SD
EIT scores	30	103	70.35	19.54
OPT scores	50	103	75.13	12.94
Prof mean scores	44	103	72.87	14.74
BWDS scores	3	8	4.25	.968
SNRT scores	10	22	15.62	2.93
WMC mean scores	5.36	11.72	8.31	1.48
LoR (in months)	1	84	18.67	23.05

7.3.3.Preliminary screening of data for statistical analysis

Prior to carrying out any inferential statistics, the assumptions for violations of normality, linearity and homogeneity were checked. A Shapiro-Wilk test was used in order to test whether the dependent variables of L1 and L2 fluency measures could be assumed to be normally distributed. The results for this statistic indicated that the data were normally distributed for all the variables with $p > .05$, except for two measures: Turkish total repair and Turkish articulation rate (Sig. value of less than .05). Although the Sig. values for these two measures suggested a violation of normality, this could be quite common in large data sets (Pallant, 2013), and therefore, an inspection of visuals (i.e. normal probability plots labelled as Q-Q plots) provided a better picture of the sample in terms of the normality. Q-Q plots obtained for the fluency measures in L1 and L2 are presented below in figures between Figure 7.1to Figure 7.10.

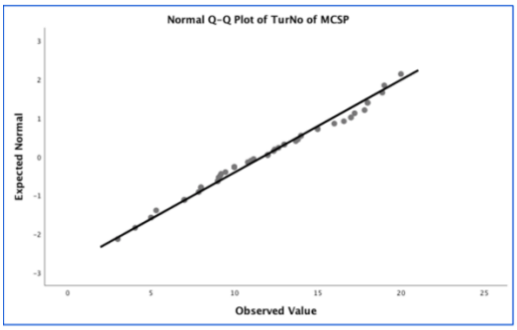


Figure 7.1. Q-Q plot for Turkish number of MCSP

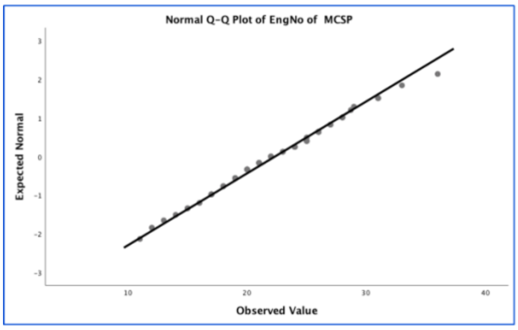


Figure 7.2. Q-Q plot for English number of MCSP

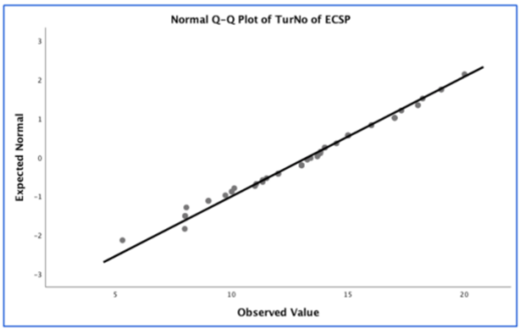


Figure 7.3. Q-Q plot for Turkish number of ECSP

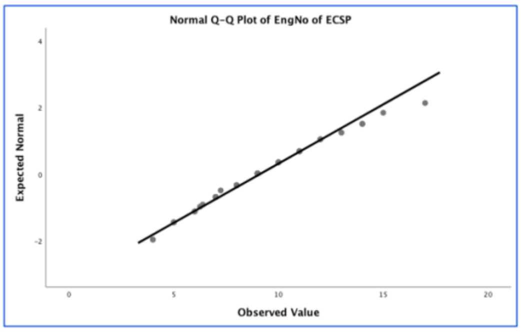


Figure 7.4. Q-Q plot for English number of ECSP

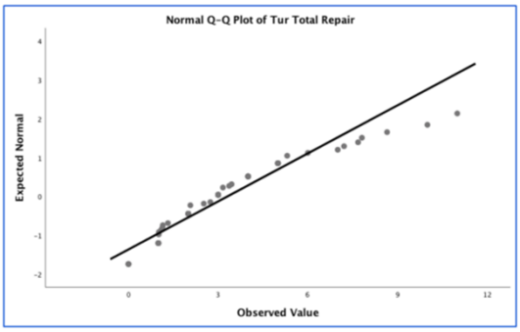


Figure 7.5. Q-Q plot for Turkish total repair

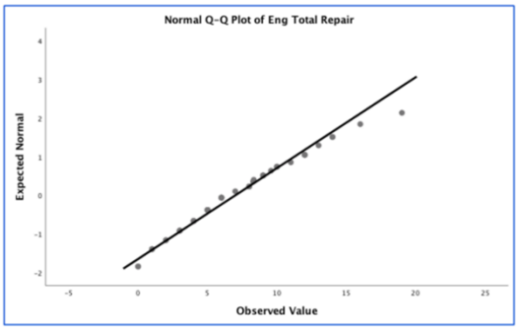


Figure 7.6. Q-Q plot for English total repair

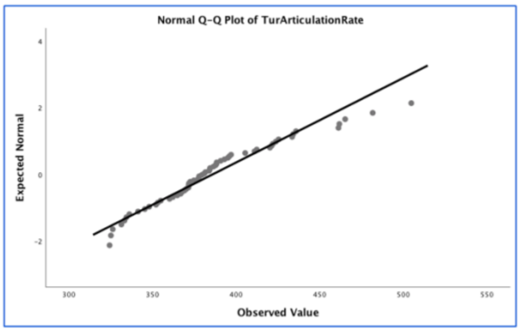


Figure 7.7. Q-Q plot for Turkish articulation rate

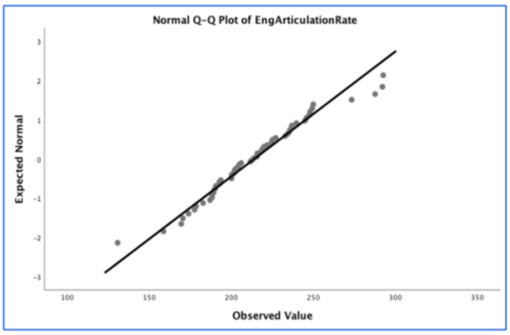


Figure 7.8. Q-Q plot for English articulation rate

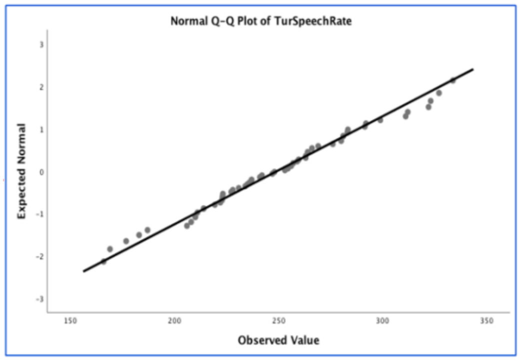


Figure 7.9. Q-Q plot for Turkish speech rate

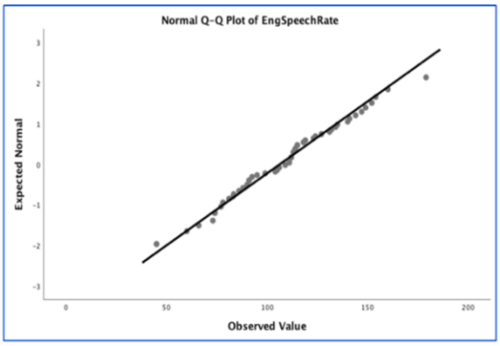


Figure 7.10. Q-Q plot for English speech rate

As can be seen from the figures above, Q-Q plots for the fluency measures in Turkish and English indicated reasonably straight lines, suggesting that in general, the dependent variables (i.e. L1 and L2 fluency measures) were approximately normally distributed. Some slight skewedness was also observed for total repair in both Turkish and English, however given that the skewedness was very slight in both, no action was taken to transform the data.

Violations of assumptions of normality were also checked for the independent variables of the study, i.e. L2 proficiency (OPT, EIT and the mean score), WMC (BWDS, SNRT and the mean score) and LoR. Shapiro-Wilk test was carried out for each independent variable, and the results indicated that the data were normally distributed with $p > .05$. An inspection of Q-Q plots of these variables for Pearson correlational analysis (Figure 7.11 to Figure 7.17 below) further supported that there were no violations of assumptions of normality.

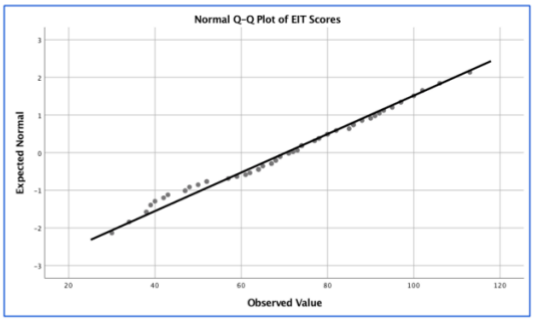


Figure 7.11. Q-Q plot for EIT scores

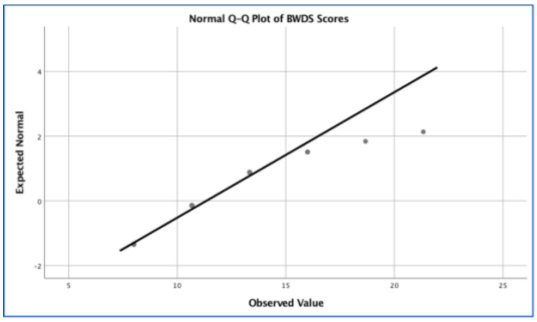


Figure 7.13. Q-Q plot for BWDS scores

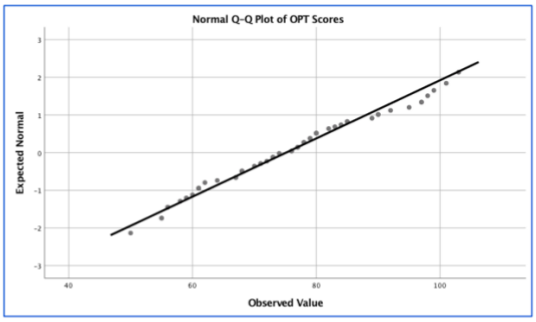


Figure 7.12. Q-Q plot for OPT Scores

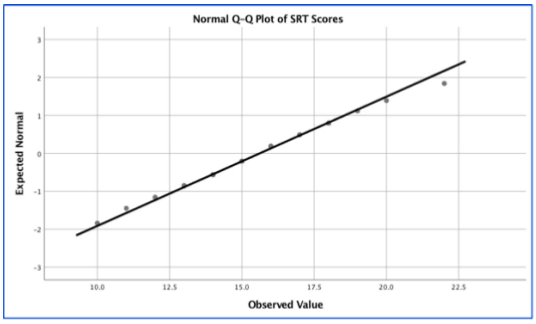


Figure 7.14. Q-Q plot for SNRT scores

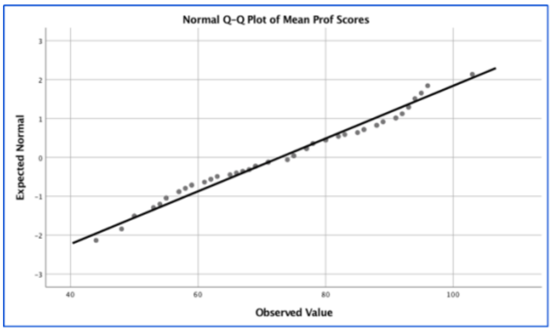


Figure 7.15. Q-Q plot for mean scores of L2 proficiency

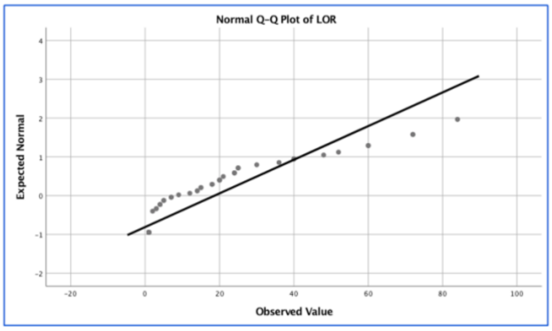


Figure 7.17. Q-Q plot for LoR

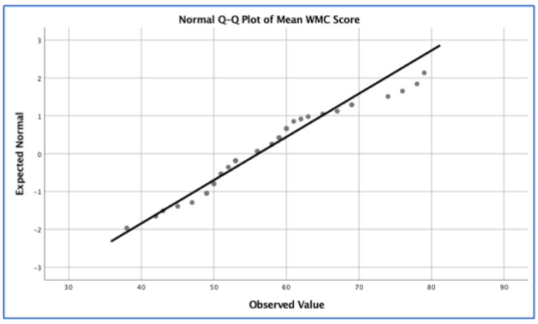


Figure 7.16. Q-Q plot for mean scores of WMC

Further, the assumptions of linearity and homoscedasticity were checked; scatterplots were obtained for each fluency measure in the two languages. It is recommended that scatterplots are obtained before carrying out correlations because if variables in question are related in a curvilinear (i.e. non-linear) fashion, the correlational analysis will seriously underestimate the strength of the relationship between the variables (Pallant, 2013, p. 66). Scatterplots also provide a picture of homoscedasticity of variance. Please see below the scatterplots (Figure 7.18 to Figure 7.22) obtained for each fluency measure; the horizontal axis in each figure indicates English fluency measure while the vertical axis indicates Turkish fluency measure.

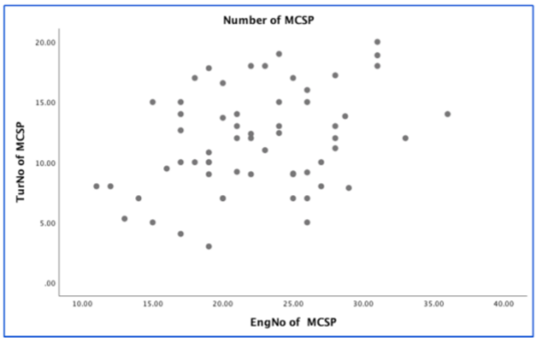


Figure 7.18. Scatterplot for number of MCSP

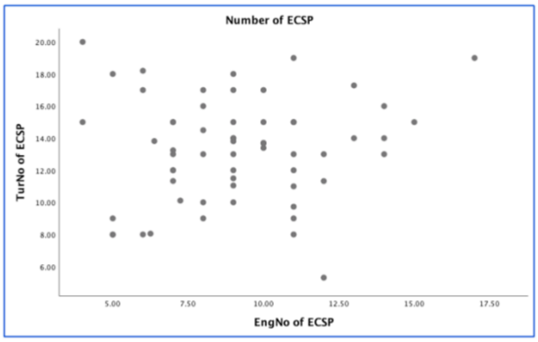


Figure 7.19. Scatterplot for number of ECSP

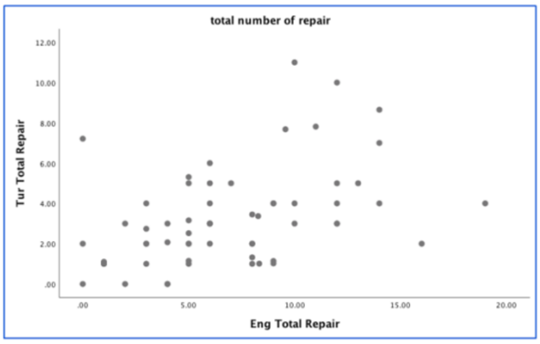


Figure 7.20. Scatterplot for total repair

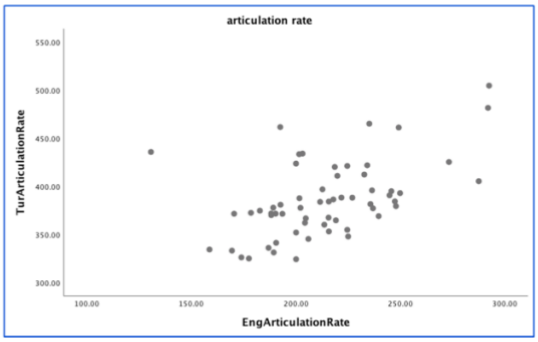


Figure 7.21. Scatterplot for articulation rate

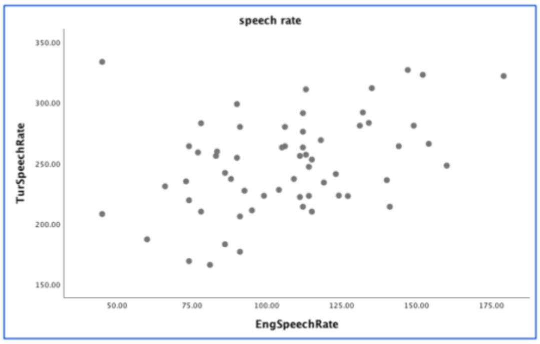


Figure 7.22. Scatterplot for speech rate

As can be seen from the figures above, the relationships between the variables were roughly linear and the scores were approximately evenly spread, suggesting no violation of the assumptions of linearity and homoscedasticity. Therefore, it was appropriate to carry out Pearson product-moment correlations to explore the relationship between fluency measures in both languages.

7.3.4.Pearson product-moment correlations

The RQ1 examined whether there was any relationship between L1 and L2 measures. The relationships between the two languages were investigated using Pearson product-moment correlation coefficient for all fluency aspects (i.e. breakdown, repair, speed and composite) separately. Similar to Study 1, both variables in the current study, i.e. L1 and L2 measures, were continuous variables (interval level); therefore, a simple bivariate correlation was deemed appropriate to answer the RQ. Further, correlational analysis would provide information about the strength and the direction of linear relationships, if any (Lowie & Seton, 2013; Pallant, 2013). Results of the Pearson correlations are presented in Table 7.3. Please note that the sizes of the value of Pearson correlations were interpreted using the field-specific criteria recommended in Plonsky and Oswald (2014); i.e. benchmarks of $r = .25$ (small), $r = .40$ (medium) and $r = .60$ (large).

It can be seen from the table that moderate to strong positive correlations were observed for all measures, except *number of end-clause silent pauses*. As regards to breakdown fluency measures, while there was a moderate positive correlation for *number of mid-clause silent pauses* between the languages ($r = .35$, $p = .006$), no correlation was observed for n*umber of end-clause silent pauses* ($r = .11$, $p = .391$). Regarding the other aspects of fluency, there were moderate to strong correlations for *total number of repair* ($r = .44$, $p = .001$), *articulation rate* ($r = .51$, $p = .001$) and *speech rate* ($r = .39$, $p = .002$). These results suggest a strong relationship between L1 and L2 fluency measures.

Table 7.3. Correlations between L1 and L2 fluency measures for all groups in Study 2

Fluency measures	*r*	*p*
	Breakdown fluency	
Number of MCSP	.351**	.006
Number of ECSP	.113	.391
	Repair fluency	
Total repair	.440**	.001
	Speed fluency	
Articulation rate	.511**	.001
	Composite measures	
Speech rate	.391**	.002

**Correlation is significant at the 0.05 level (2-tailed)*

***Correlation is significant at the 0.01 level (2-tailed)*

Further to the simple correlations, coefficients of determination for the significant results were calculated in order to have an idea of how much variance was shared by the measures in L1 and L2. Similar procedures in Study 1 were followed (Pallant, 2013); the *r* values were multiplied by themselves and a 'percentage of variance' was obtained by further multiplying by 100 (e.g. for a correlation at r = .578, it should be (.578 x .578) x 100 = 33.40 percent of variance). These squared values indicate how much of the variance (in percentage) in participants' L2 fluency behaviour could be explained by their L1 fluency behaviour. Table 7.4 illustrates R squared values for the significant correlations observed.

Table 7.4. Coefficients of determination for the significant results in Study 2

Fluency measures	R squared
Number of MCSP	12.32
Total repair	19.36
Articulation rate	26.11
Speech rate	15.28

As can be seen in the table, the amount of variance shared by L1 and L2 fluency measures was 12.32% for *number of mid-clause silent pauses*, 19.36% for *total repair*, 26.11 % for *articulation rate* and 15.28 % for *speech rate*. Following from Plonsky and Ghanbar (2018), who proposed the benchmarks of .20 (or below) and .50 (or above) as small and large, respectively, these figures could be considered respectable amounts of variance in English fluency behaviour explained by Turkish fluency behaviour.

It was also explored whether the significant correlations found in the Pearson correlations were similar across different proficiency groups. To investigate the relationships between Turkish and English fluency measures across the groups, the data file was first split and then the analysis was repeated. Table 7.5 below presents the results from Pearson correlations across different proficiency levels (i.e. B1, B2 and C1 levels).

Table 7.5. Correlations between L1 and L2 fluency measures across different proficiency levels in Study 2

Fluency measures	**B1 (n=18)**		**B2 (n=21**		**C1 (n=21)**	
	r	***p***	***r***	***p***	***r***	***p***
Breakdown fluency						
Number of MCSP	.316	.201	.396	.076	.419	.059
Number of ECSP	.378	.122	-.239	.296	.176	.445
Repair fluency						
Total repair	508*	.031	.553*	.009	.242	.290
Speed fluency						
Articulation rate	.316	.202	.686**	.001	.579**	.006
Composite measures						
Speech rate	.247	.323	.509*	.018	.675**	.001

**Correlation is significant at the 0.05 level (2-tailed)*

***Correlation is significant at the 0.01 level (2-tailed)*

As can be seen in the table, there were significant correlations for all proficiency groups, i.e. B1, B2 and C1, between L1 and L2 fluency measures. Strong positive correlations were observed in B2 and C1 levels for *articulation rate* (r = .68, p = .001 and r = .57, p = .006, respectively) and for *speech rate* (r = .50, p = .018 and r = .67, p = .001, respectively). No significant correlation was found for B1 level for these two measures.

As for *total repair*, while there were strong positive correlations for B1 and B2 levels ($r = .50$, $p = .031$ and $r = .55$, $p < .009$, respectively), no correlation was observed for C1 level. Interestingly, the results did not indicate any correlations for breakdown measures when the proficiency groups were analysed separately. Related to this, however, it is worth mentioning that C1 level just missed the significance level for *number of mid-clause silent pauses* ($r = .41$, $p = .059$).

7.3.5. Partial correlations

RQs 1(a), 1(b) and 1(c) examined the extent to which the relationships between L1 and L2 fluency measures were moderated by individual factors of L2 proficiency (i.e. mean scores), WMC (i.e. mean scores) and LoR, respectively. In order to address these questions, partial correlations were carried out separately with each of these independent variables as the controlling factor. Partial correlations were employed in Study 2 because this analysis allowed the researcher to statistically control for these individual variables by removing their influence, and thus to have a clearer picture of the relationship between the fluency measures in L1 and L2. In addition, this analysis was used in Study1, too, to control for the effect of L2 proficiency and it worked well for this purpose. Therefore, using partial correlations, we could understand to what extent the relationships between the two languages are influenced by the independent variables of Study 2 (i.e. L2 proficiency, WMC and LoR). If there is a big influence from one of these variables, the zero-order correlations between the two languages (i.e. significant correlations from Pearson correlations in the previous section) would decrease, suggesting a weaker relationship between the variables.

7.3.5.1. Partial correlations controlling for L2 proficiency

In the previous section, simple bivariate correlations were carried out across the proficiency levels (i.e. B1, B2 and C1 levels) in order to see whether the correlations observed between L1 and L2 fluency measures were similar in different proficiency groups. In this section, the significant correlation coefficients between fluency measures in the two languages were further explored by running partial correlations, with language proficiency (i.e. mean score of L2 proficiency) as the controlling factor, in order to explore whether L2 proficiency mediates these relationships. The results from the partial correlations controlling for L2 proficiency are provided in Table 7.6 below.

As can also be seen in the table, the results indicated that moderate to strong positive correlations between Turkish and English for all fluency measures were maintained, with high numbers of these measures in English being associated with high numbers in Turkish. An inspection of the zero order correlations suggested that controlling for L2 proficiency had little impact on the strength of the relationship for *number of mid-clause silent pauses* (r = .35, p < .006), *total repair* (r = .44, p < .001) and *articulation rate* (r = .52, p < .001) (zero order correlations r = .35, p < .006, r = .44, p < .001 and r = .51, p<.001, respectively). For *speech rate*, there was a small increase in the strength of the correlation (from .39 to .43), suggesting that L2 proficiency had a very little influence on the strength of the relationship between L1 and L2 speech rate. The results, when taken together, suggest that the observed relationships between L1 and L2 fluency measures are not due to the influence of language proficiency.

Table 7.6. Partial correlations for the significant results, controlling for language proficiency in Study 2

Fluency measures	***r***	***p***	**Zero order (Pearson) correlation**
Number of MCSP	.355	.006	.351, p<.006
Total repair	.440	.001	.440, p<.001
Articulation rate	.522	.001	.511, p<.001
Speech rate	.435	.001	.391, p<.002

7.3.5.2. Partial correlations controlling for working memory capacity

In order to explore whether the relationships between L1 and L2 fluency behaviours are mediated by WMC, partial correlations for the significant correlation coefficients from the Pearson correlational analysis were carried out, with WMC (i.e. mean score of WMC) as the controlling factor. The results demonstrated that this variable had very little impact on the strength of the relationships between L1 and L2 fluency measures. When partial correlations were compared with zero order correlations, it can be seen that the significant results were maintained for *number of mid-clause silent pauses* at r = .35, p < .006 and *total repair* at r = .44, p < .001, with zero order correlations of r = .35, p < .006 and r = .44, p < .001, respectively. There was a small decrease in the strength of the correlation for *articulation rate* (from .51 to 46) and *speech rate* (from .39 to .37), suggesting that WMC had very little influence on the strength of the relationships between L1 and L2 fluency behaviours for these measures. These results overall suggest that the observed relationships between L1 and L2 fluency measures were not due to the effect of WMC. Table 7.7 below illustrates the results obtained from partial correlations controlling for this variable.

Table 7.7. Partial correlations for the significant results, controlling for WMC in Study 2.

Fluency measures	***r***	***p***	**Zero order (Pearson) correlation**
Number of MCSP	.352	.006	.351, p<.006
Total repair	.442	.001	.440, p<.001
Articulation rate	.467	.001	.511, p<.001
Speech rate	.375	.003	.391, p<.002

7.3.5.3. Partial correlations controlling for length of residence

Further partial correlations were run once again for the significant correlation coefficients from Pearson correlational analysis, this time controlling for LoR as the controlling variable, in order to explore whether LoR mediates the relationships between L1 and L2 fluency behaviour. Similar to the results from the partial correlations controlling for the other two variables (i.e. L2 proficiency and WMC), the results for this variable, too, demonstrated that the significant correlations were maintained for all fluency measures. A inspection of the zero order correlations suggested that overall the amount of time that the participants spent abroad had very little influence on the strength of the relationships for *number of mid-clause silent pauses*, *total repair* and *articulation rate* (r = .35, p <. 005, r =.42, p < .001 and r = .51, p < .001, respectively (zero order correlations of r = .35,

$p < .006$, $r = .42$, $p < .001$ and $r = .51$, $p < .001$, respectively). Similar to the results from the partial correlations controlling for L2 proficiency, a small increase was observed in the strength of the correlation for *speech rate* (from .39 to .45), suggesting that LoR had very little impact on the relationship between Turkish and English speech rate. The results from partial correlations controlling for LoR are provided in Table 7.8 below.

Table 7.8. Partial correlations for the significant results, controlling for LoR in Study 2

Fluency measures	*r*	*p*	Zero order (Pearson) correlation
Number of MCSP	.359	.005	.351, p<.006
Total repair	.421	.001	.440, p<.001
Articulation rate	.517	.001	.511, p<.001
Speech rate	.457	.001	.391, p<.001

In sum, the partial correlations controlling for the three independent variables of Study 2 (i.e. L2 proficiency, WMC and LoR) did not demonstrate any influence of these variables on the strength of the relationships between fluency measures in L1 and L2. In other words, it could be suggested that the relationships between Turkish and English fluency behaviour are not due to the effect of these particular variables.

7.4. Analysis of lexical measures

7.4.1. Preliminary analysis for lexical complexity (LD) measures

Before running any statistical analysis for LD measures, the data was checked for possible errors, i.e. those which might have occurred in the data entry, as well as for any outliers. Since no errors or outliers were detected at this stage, descriptive statistics were run to have a first impression about the sample and to explore the differences between L1 and L2 LD measures. Further to the descriptive statistics, paired samples t-tests were run in order to see if the differences between the LD measures in the two languages were statistically significant. Cohen's effect sizes were also calculated, and these were interpreted using the field-specific criteria recommended in Plonsky and Oswald (2014).

Table 7.9 below presents the descriptive statistics of LD measures, results from t-tests and the effects sizes.

As can be understood from the table, the LD values seemed different between Turkish and English for TTR values, with *M* = .56, *SD* = .09 and *M* = .44, *SD* = .07, respectively and for D scores with *M* = 48.82, *SD* = 19.07 and *M* = 23.09, *SD* = 9.34, respectively. The results from the t-tests further indicated that the differences between these measures, i.e. TTR values and D scores, in the two languages were statistically significant, with $t = 9.48$, $p < .001$ and $t = 10.39$, $p < .001$, respectively. Cohen's effect sizes were large with d = 1.48 for TTR values and d = 1.71 for D scores (Plonsky & Oswald, 2014). Taken together, these statistics suggest that the participants performed differently in both languages when their lexical complexity behaviour was measured using simple *TTR values* and *D measure* and the differences were important.

Table 7.9. Descriptive statistics for LD measures in L1 and L2 in Study 2.

D measures	Mean	SD	*t*	Sig (2-tailed)	CIs (for Mean)		Cohen's *d*
Turkish TTR	.56	.09	9.48	.001	Lower Bound	.53	1.48
					Upper Bound	.58	
English TTR	.44	.07			Lower Bound	.42	
					Upper Bound	.46	
Turkish D	48.82	19.07	10.39	.001	Lower Bound	43.89	1.71
					Upper Bound	53.74	
English D	23.09	9.34			Lower Bound	20.68	
					Upper Bound	25.51	

7.4.2.Preliminary screening of data for statistical analysis

Before carrying out any inferential statistics for LD measures, violations of assumptions for normality, linearity and homoscedasticity were checked. Shapiro-Wilk test was used in order to test whether the TTR values and D scores could be assumed to be normally distributed. The results indicated that the data was normally distributed with $p > .05$. An inspection of Q-Q plots also supported the results of this test, showing that the data was approximately normally distributed. (Please see Figure 7.23 to Figure 7.26 below for Q-Q plots for TTR values and D scores in Turkish and English.

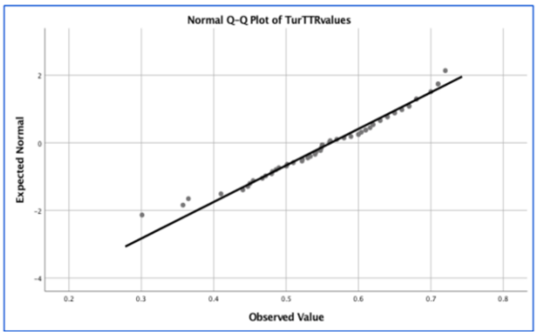


Figure 7.23. Q-Q plot for Turkish TTR values

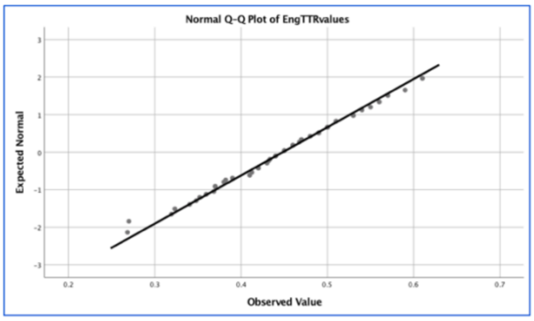


Figure 7.24. Q-Q plot for English TTR values

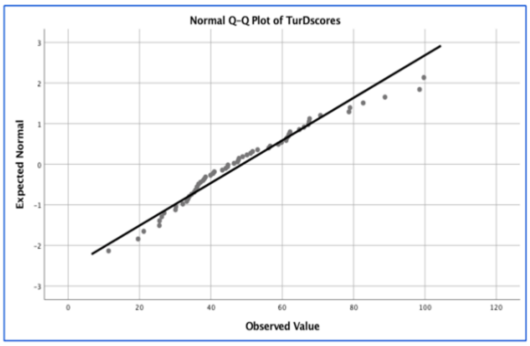


Figure 7.25. Q-Q plot for Turkish D scores

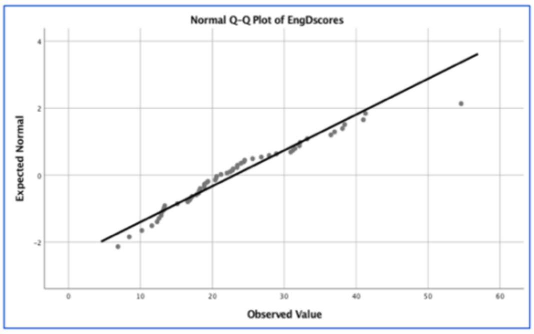


Figure 7.26. Q-Q plot for English D scores

In order to check for violations of assumptions for linearity and homoscedasticity, scatterplots were obtained for each LD measures in both languages, which are presented below in Figure 7.27-Figure 7.28.

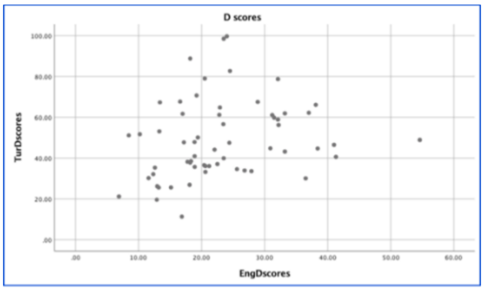


Figure 7.27. Scatterplot for D scores

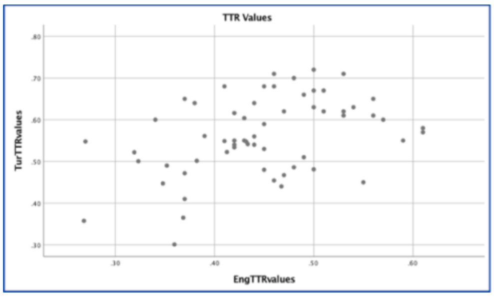


Figure 7.28. Scatterplot for TTR values

7.4.3. Pearson product-moment correlations for lexical complexity (LD) measures in L1 and L2

The second RQ of Study 2 examined the relationship between L1 and L2 lexical complexity represented through calculations of *TTR values* and *D measure* (i.e. D scores) in both languages. In order to address the question, simple bivariate correlations, i.e. Pearson product-moment correlations, were run between TTR values and D scores in both Turkish and English. In the same line of reasoning for the analysis of fluency measures, Pearson correlations were deemed appropriate to answer this RQ as both variables under examination, i.e. TTR values and D scores, were continuous (i.e. interval level). Further, correlational analyses would reveal not only the relationship between L1 and L2 lexical measures, if any, but also would describe the strength and the direction of the relationship. The results from Pearson correlations for TTR values and D scores in L1 and L2 are provided in Table 7.10 below. Please note that the size of the effect of the correlations were interpreted following the recommendations in Plonsky and Oswald (2014), i.e. benchmarks of r = .25 (small), r = .40 (medium) and r = .60 (large).

Table 7.10. Correlations between L1 and L2 for TTR values and D scores in Study 2

LD measures	*r*	*P*
TTR values (traditional type-token ratio)	.41**	.001
D measure (D scores)	.23	.072

**Correlation is significant at the 0.05 level (2-tailed)*

***Correlation is significant at the 0.01 level (2-tailed)*

As can be seen in Table 7.10, the results demonstrated a moderate positive correlation between TTR values in Turkish and English (r = .41, p = .001), while no correlation was observed between the two languages for D scores. In order to obtain the coefficients of determination for the significant results, i.e. TTR values, the procedures described in Pallant (2013) were followed once again. For this, the r values were multiplied by themselves first and then they were further multiplied by 100 to obtain a percentage value (i.e. r Table 7.11). This figure suggests that Turkish TTR values helps to explain nearly 17 percent of the variance in English TTR values, which could be considered a small amount of variance (Plonsky & Ghanbar, 2018).

Table 7.11. Coefficients of determination for the significant results in Study 2

Lexical Complexity measures	R squared
TTR	16.81

7.4.4. Pearson product-moment correlations for fluency and lexical complexity (LD) measures in each language

The third RQ of Study 2 sought to examine the link between fluency behaviour and lexical complexity within each language, if any (i.e. relationship between L1 fluency behaviour and L1 lexical complexity, and between L2 fluency behaviour and L2 lexical complexity). Similar to the analyses run to answer the second RQ, Pearson product-moment correlational analysis was used to address this RQ. Simple bivariate correlations would demonstrate if any of the fluency measures in one language (i.e. Turkish and English) would correlate with the lexical complexity values in the same language. After ensuring that normality, linearity and homoscedasticity were not violated (please see section 7.3.3 above), simple correlations were carried out between fluency and LD measures in each language separately (e.g. correlations between Turkish fluency measures and Turkish LD measures).Table 7.12 provides the results obtained from the correlational analysis between fluency and LD measures in *Turkish*.

Table 7.12. Correlations between fluency measures and LD measures in Turkish in Study 2

	TTR values		**D Scores**	
Fluency measures in Turkish	***r***	***p***	***r***	***p***
Breakdown fluency				
Number of MCSP	-.254	.050	-.230	.077
Number of ECSP	.025	.853	.037	.779
Repair fluency				
Total repair	.006.	.964	-.069	.598
Speed fluency				
Articulation rate	.287*	.026	.203	.120
Composite measures				
Speech rate	.391**	.002	.270*	.037

**Correlation is significant at the 0.05 level (2-tailed)*

***Correlation is significant at the 0.01 level (2-tailed)*

As can be seen in the table, no significant correlation was found between breakdown and repair fluency measures (i.e. for *number of mid-clause silent pauses*, *number of end-cause silent pauses* and *total repair)* and LD measures in Turkish. As for speed and composite measures, weak to moderate positive correlations were observed between *TTR values* and *articulation rate* (r = .28, p < .026) and between *TTR values* and *speech rate* (r = .39, p < .002). While there was a weak positive correlation between *D scores* and *speech rate* (r = .27, p < .037), no correlation was found between *D scores* and *articulation rate*. Taken together, the results suggest that as far as Turkish as L1 is concerned, fluency and lexical complexity (represented through LD) aspects of oral performance could be linked to each other to a small extent.

When the same statistical analysis was carried out to explore the relationship between fluency and LD measures in *L2 English* (Table 7.13), the statistics demonstrated similar results to those observed for Turkish measures; no statistically significant correlation was found between LD measures (i.e. TTR values and D scores) and fluency measures of breakdown and repair, i.e. *number mid-clause silent pauses*, *number of end-clause silent pauses* and *total repair*. When we turn to the speed and composite fluency measures, however, the figures indicate that both TTR values and D scores correlated significantly with *speech rate*. There was a moderate positive correlation between *speech rate* and *TTR values* at r = .39, p = .002, and between *speech rate* and *D scores* at r = .35, p = .005. It is interesting to see that neither of the LD measures correlated with articulation rate in L2 whereas TTR values were found to correlate with articulation rate in L1. Overall, these results suggest that fluency and lexical complexity in L2 English could be related to each other to a small extent.

Table 7.13. Correlations between fluency measures and LD measures in English in Study 2

	TTR values		**D scores**	
Fluency measures in English	***r***	***p***	***r***	***p***
Breakdown fluency				
Number of MCSP	-.159	.224	-.207	.113
Number of ECSP	.115	.381	.148	.258
Repair fluency				
Total repair	.145	.267	.158	.227
Speed fluency				
Articulation rate	.149	.256	.147	.263
Composite measures				
Speech Rate	.394**	.002	.358**	.005

**Correlation is significant at the 0.05 level (2-tailed)*

***Correlation is significant at the 0.01 level (2-tailed)*

Further to these simple correlations between fluency and LD measures in the two languages, it was explored whether the significant correlations, which were yielded in each language persisted across different proficiency groups (Table 7.14). Only speech rate was explored in this analysis because it was the only fluency measure for which significant correlations were observed in both languages. Pearson correlations were carried out once more to analyse each group separately after the data file was split. The results demonstrated that for TTR values, significant strong correlations were observed for the highest proficiency group (i.e. C1 level) between *speech rate* and *TTR values* in both Turkish and English ($r = .59$, $p < .004$ and $r = .46$, $p = .033$, respectively). Turkish *speech rate* also correlated with Turkish *TTR values* in B2 level at $r = .43$, $p < .048$. As for the correlations between D scores and speech rate in both languages, there was a moderate strong correlation for C1 level in Turkish ($r = .45$, $p < .037$, however, the correlation in English for these measures just missed the significance level in this group ($r = .42$, $p < .053$). The results overall suggest that the fluency measure of speech rate and LD measures (i.e. TTR values and D scores) correlate with each other at higher levels of L2 English.

Table 7.14. Correlations between speech rate and LD measures in L1 and L2 across different proficiency levels in Study 2

	B1 (n=18)				B2 (n=21)				C1 (n=21)			
	Turkish		English		Turkish		English		Turkish		English	
	r	***P***	***r***	***p***	***r***	***P***	***r***	***p***	***r***	***p***	***r***	***P***
TTR	.266	.286	.213	.397	.437*	.048	.305	.179	.596**	.004	.467*	.033
D scores	.156	.536	.179	.477	.267	.242	.197	.391	.458*	.037	.428	.053

**Correlation is significant at the 0.05 level (2-tailed)*

***Correlation is significant at the 0.01 level (2-tailed)*

7.5. Summary of chapter and the key findings

In this chapter, I have presented the results from the statistical analysis carried out to address the RQs of Study 2. The findings were presented under two main sections: analysis of fluency measures and analysis of lexical complexity (i.e. LD) measures. Each section started with the preliminary analysis, i.e. descriptive statistics and the visuals for the assumptions of the statistical analysis. Then, I moved on to present the findings from the inferential analysis run to answer each RQ. The three RQs that Study 2 aimed to address, and the summary of the findings in relation to these are provided in the table below (Table 7.15).

Table 7.15. The research questions and an overview of the key findings in Study 2

	Research questions	Overall findings
RQ1-	To what extent L1 Turkish and L2 English fluency behaviours are related among higher-proficiency learners in a study-abroad context?	To a certain extent. Significant correlations were observed between L1 and L2 fluency measures for all aspects of fluency. Moderate to strong positive correlations found between L1 and L2 fluency measures: - number of mid-clause silent pauses (*r*=.35) - total number of repair (*r*=.44) - articulation rate (*r* =. 51) - speech rate (*r*= .39)
RQ1a)	Does level of L2 proficiency, measured through OPT and EIT, mediate the relationship between L1 and L2 fluency behaviours among high-proficiency learners in a study-abroad context?	Overall, no impact of L2 proficiency was observed on the strength of the relationships between L1 and L2 fluency measures.

RQ1b)	Does working memory capacity (WMC), measured through BWDS and SNRT, mediate the relationship between L1 and L2 fluency behaviour among high-proficiency learners in a study-abroad context?	Overall, no impact of WMC was observed on the strength of the relationships between L1 and L2 fluency measures.
RQ1c)	Does length of residence (LoR) mediate the relationship between L1 and L2 fluency behaviours among high-proficiency learners in a study-abroad context?	Overall, no impact of LoR was observed on the strength of the relationships between L1 and L2 fluency measures.
RQ2)	To what extent are L1 Turkish and L2 English lexical complexity behaviours, represented through LD scores, related among higher-proficiency learners in a study-abroad context?	To a small extent. For TTR values only, a significant moderate correlation was observed between L1 and L2 LD scores (*r*= .41)

RQ3)	To what extent are fluency and lexical complexity aspects of second language oral performance for both languages (Turkish and English) related among higher-proficiency learners in a study-abroad context??	To a small extent. In both L1 Turkish and L2 English, significant correlations were observed between fluency measures and LD measures: Weak to moderate correlations found between fluency and LD measures in Turkish: - TTR values and articulation rate (*r* = .28) - TTR values and speech rate (*r* = .39) - D scores and speech rate (*r* = .27) Moderate correlations found between fluency and LD measures in English: - TTR values and speech rate (*r* = .39) - D scores and speech rate (*r* = .35)

CHAPTER 8. DISCUSSION: STUDY 2

8.1. Introduction

This chapter interprets and discuss the findings of Study 2 in the light of the relevant literature and previous studies. It is organized into three main sections; I start with a brief overview of the findings from Study 2 and discuss these under two further sub-sections of fluency and lexical complexity. Under the fluency section, the discussion centres on the findings in relation to 1) overall differences between L1 and L2 fluency, 2) correlations observed between the two in general and across different proficiency levels, and 3) partial correlations for significant results controlling for learner individual factors, i.e. LP, WMC and LoR. The second section discusses the findings relating to the lexical complexity aspect and covers 1) the differences between L1 and L2 LD scores, 2) correlations between them and 3) correlations between fluency and lexical complexity measures in each language. The final section concludes the chapter with a summary of the key points discussed throughout.

8.2. Overview of the findings in study 2

Study 2 aimed to examine the relationship between L1 and L2 oral performance with respect to fluency and lexical complexity aspects of performance for learners of higher-proficiency levels (i.e. B1, B2 and C1) in a study-abroad context. Regarding the fluency aspect, this goal originated from the results of previous studies which suggested that speakers' personal speaking styles might be reflected in their L2 speech fluency (e.g. De Jong et al., 2015; Huensch & Tracy-Ventura, 2016; Peltonen, 2018). An important issue raised by the research was with regard to trait versus language-specific state, i.e. to what extent fluency is a characteristic specific to an individual or a certain language being spoken. It was therefore suggested that sources of variability that are not L2-specific but influence judgements of speakers' L2 fluency (e.g. personal styles) should be removed from L2 speech so that only L2-specific aspects are considered (De Jong et al., 2015; Segalowitz, 2010). Therefore, an investigation of the relationship between L1 and L2 fluency behaviour was deemed crucial.

While Study 1 was aimed at exploring this relationship for learners of low-proficiency levels in an EFL context, Study 2 attempted to replicate and expand the findings of Study 1 by further examining this relationship for learners of higher-proficiency levels in a study-abroad context. To this end, speech data in L1 and L2 were collected from a group of 60 Turkish post-graduate students in the UK, and were analysed for five fluency measures to represent all fluency aspects: two breakdown measures (i.e. number of mid-clause silent pauses and number of end-clause silent pauses), one repair measure (i.e. total number of repair instances), one speed fluency measure (i.e. articulation rate) and one composite measure (i.e. speech rate, which incorporated breakdown and speed aspects of fluency). The participants' proficiency levels were measured through two standardized tests, i.e. OPT and EIT. A further aim of Study 2 was to explore whether variations in individual learner variables, i.e. LP, WMC and LoR, mediated factors in determining the strength of the relationships between L1 and L2 fluency behaviour. To achieve these aims, one research question with three sub-questions was formed and answered in the previous chapter. The results indicated that all aspects of L2 fluency, i.e. breakdown, repair and speed, were related to L1 fluency behaviour, at least to some extent. The results also revealed that overall L1-L2 relationships persisted regardless of the variability in individual factors, i.e. LP, WMC or LoR.

A secondary aim of Study 2 was to examine the relationship between L1 and L2 lexical complexity (represented through LD aspect only) and between fluency and lexical complexity in each language separately. The results demonstrated that L1 and L2 lexis were related to each other, to a small extent, only when the participants' lexical performance was measured using the traditional measure of TTR. The only fluency measure that correlated with TTR scores in both L1 and L2 separately was speech rate. In addition to this, only in Turkish, articulation rate and TTR scores correlated positively. In the following sections, I discuss these findings in relation to how they relate to those from earlier studies and Study 1, and I attempt to offer some explanations for them.

8.3. Fluency in L1 and L2

In this section, I discuss the results of Study 2 as they relate to the RQ1. Starting with overall differences observed between L1 and L2 fluency measures. I discuss the findings from the correlational analysis and compare these with the findings of Study 1 and other studies reported in the literature.

8.3.1. Differences in L1 and L2 fluency

The descriptive statistics highlighted a number of differences between L1 and L2 fluency behaviour. The participants exhibited large differences in their pausing behaviour for number of mid-clause and end-clause silent pauses. They were found to make more mid-clause pauses in their L2 and more end-clause pauses in their L1. This corroborates the findings of Study 1 and earlier studies (De Jong, 2016; Riazantseva, 2001; Tavakoli, 2011; Skehan, 2014) in that L2 speakers tend to pause more often within clauses and less so between clauses as compared to L1 speakers. Previous research had already shown that mid-clause pausing could especially be indicative of processing difficulties encountered in L2 speech production as L2 speech is not yet automatic (Kormos, 2006). The descriptive statistics of the present study seem to lend further support to previous studies in this regard.

The participants also differed in their L1 and L2 with respect to their repair and speed fluency behaviours, and the results of the t-tests indicated that these differences were statistically significant. The higher figures for repair fluency in English and for speed fluency in Turkish imply that the participants were overall less fluent in their L2 than their L1 speech, also supporting the findings of previous studies (De Jong et al., 2016; Huensch & Tracy-Ventura, 2016)

8.3.2. Correlations between L1 and L2 fluency

The first RQ examined whether there was a relationship between L1 and L2 fluency behaviour for higher-proficiency learners in a study-abroad context. It was hypothesized that the stronger relationships between L1 and L2 fluency measures would mean that L2 fluency measures reflect more personal styles of the participants in their L1. In answering this question, it was found that L1 and L2 fluency behaviours were linked to each other, at least to some extent, for all fluency aspects, confirming the findings of previous studies (De Jong et al., 2015; Huensch & Tracy-Ventura, 2016; Peltonen, 2018). The correlational analyses showed that all fluency measures in L1 and L2, except for number of end-clause silent pauses, were positively correlated with each other although the degrees of the strength varied.

8.3.2.1. Breakdown fluency

With respect to breakdown fluency, the silent pauses were examined in their frequency and location. The significant positive correlation for the breakdown measure of mid-clause silent pauses (r = .35) between the two languages suggested that the participants' L2 mid-clause pausing was to some extent a function of their L1 mid-clause pausing. This result is consistent with the findings of Study 1 and Peltonen (2018) in which moderate positive correlations for this measure were reported (r = .34 and r = .66, respectively). De Jong (2016) suggested that L2 mid-clause pausing behaviour is informative of L2 proficiency. Based on the findings of the current study, it could be argued that L2 pausing behaviour within clauses could be indicative of personal styles as well, at least to some extent. At this point, it is worth noting that since the current study distinguished between within and between-clause silent pauses when measuring the frequency of pauses as opposed to most other studies, which did not make such a distinction and examined only total number of silent pauses(e.g. De Jong et al., 2015; Derwing et al., 2009; Huensch & Tracy-Ventura, 2016).Therefore, it is not feasible to compare the results with those from these studies. Given the research evidence suggesting that L1 and L2 speakers are more likely to differ in their pause location (i.e. within or between clauses) (De Jong, 2016; Tavakoli, 2011; Skehan & Foster, 2008), this distinction between within and between-clauses is a valuable contribution of the current study.

It was also found that L1 and L2 fluency measures for number of end-clause silent pauses did not correlate with each other. This finding is somewhat surprising; it is in line with the findings of Study 1 while it contradicts with Peltonen (2018) which reported moderate correlations for this measure (r = .49). Previous research had suggested that pauses, be it silent or filled, between clauses (or AS-units) are not informative of L2 proficiency; rather, they could reflect conceptual planning (e.g. De Jong, 2016) or discourse processing (Skehan et al., 2016). This is because speakers pause at these positions to conceptualize what they are going to say next, regardless of the language they speak, i.e. L1 or L2. Following from this line of reasoning, one could expect to find a correlation between L1 and L2 end-clause pausing behaviour as the participants could be assumed to engage in the conceptualization process in similar ways (Skehan, 2015). Surprisingly, however, no correlation for this measure was observed in the current study. Two explanations could be offered for this unanticipated finding. The first one relates to the researcher's observation during the data collection and analyses phases; it was observed that when narrating the stories, the participants exhibited a tendency to pause frequently after conjunctions such as '*and*', '*and then'* or '*then*' in English or '*ve*', '*ve sonra*, '*sonra*' in Turkish as in the example excerpts below:

1. | bu sırada kedi bir tüfek alıp **::** geliyor | ve erm 0.46 (0.80) kopek de (0.77) sinirlendiği için |
2. | dog catchers catch the 0.44 the big dog | and (0.75) the mouse (0.55) erm wants **::** to help him|

Following from Foster et al. (2000), pauses after such conjunction words were identified and coded as *'mid-clause'* pauses in the participants' speech samples. This coding, however, may not have been appropriate as the majority of the participants in the current study uttered such words either immediately after they finished the previous clause or following a very brief pause (< 0.25 second), mostly when planning their next utterance. In other words, they often seemed to pause, in both L1 and L2, after uttering these conjunctions to conceptualize what they were going to say next (i.e. engaging in conceptualization-related discourse processing), rather than to look for right words or syntactic forms (i.e. engaging in formulation-linked clause processing) (Skehan et al., 2016). This seems to be in line with Kilic and Bada's (2019) findings which indicated differences between pausing patterns of L1 Turkish speakers' of L2 English and L1 English speakers, regarding two conjunctions, i.e. because and whereas. They reported that while Turkish speakers paused in English for longer durations following these conjunctions than they did preceding them, English native speakers showed an opposite pattern (i.e. pausing for longer duration before these conjunctions). When findings taken together, the post-hoc observation in the current study may also support the idea that the participants in this particular case paused more often at boundaries of idea units rather than AS-units. IUs determine the propositional complexity in a given text (oral or written) (Bulte & Housen, 2012; Vasylets, Gilabert, & Manchon, 2017), which refers to how much information is given by the speaker or writer of the text (Chafe, 1980, 1994). IUs could be suggested to represent 'the extent to which a speaker/writer encodes the ideas needed to convey a given content' (Ellis & Barkhauizen, 2005, p.154). In fact, in this sense, IUs could be informative of the degree to which the speaker engages in the conceptualization in speech production process. Although IUs could be identified based on a number of different criteria such as semantic, intonation or syntactic (see Vasylets et al., 2017 for further information), they are typically bounded by pauses in speech (though not necessarily always, as pauses can occur within IUs too) (Chafe, 1980). Taken together, it seems plausible to suggest that the participants in this case might have paused after conjunctions to convey the content in the task (e.g. to describe an event or a state) or, in other words, to engage in the conceptualization. However, since such pauses were marked as 'within-clause pauses', this might have blurred the line between mid-clause pausing and end-clause pausing in this particular case. That is to say, it is possible that the participants mid-clause pausing in L1 and L2 might be reflecting the processing involved in the conceptualization stage (which typically occur between clauses) and/or the processing difficulties emerging from other stages of the speech production process (which typically occur within clauses). If that is the case, this post-hoc observation could also partly explain the significant positive correlation found between mid-clause pausing in L1 and L2.

The second explanation for the lack of a significant correlation between L1 and L2 end-clause pausing is also based on the same observation. Given the possibility that end-clause pausing (i.e. those produced after common conjunction words) exhibited by the majority of the participants could in effect be considered as mid-clause pausing, the number of end-clause pausing could be null or non-existent in this case. It is also worth mentioning that this particular pattern observed in the participants' speech could be the result of the task type, i.e. narrative tasks. Task type has been shown to impact on fluency of performance ((Préfontaine & Kormos, 2015), and since narrating a story requires one to describe events in a logical order, this could have led the participants to sequence events using conjunction words- often automatically without actually paying much attention to them- such as '*and*' or '*and then*'. In this sense, the current study raises a methodological issue regarding the identification of the pauses; i.e. specific decisions could possibly be taken by researcher as to whether (or which) pauses after conjunction words should be regarded as within or between clause pauses, based on her observation of general tendencies emerging from participants' performance. A deeper analysis of such patterns, which occur when learners perform different task types, is likely to be another methodological issue. Further research is needed to investigate these post-hoc observations in this study.

A final observation about L1-L2 links for breakdown fluency was that when the relationship between L1 and L2 mid-clause pausing behaviour was further investigated across different LP levels, the correlations were lost for all levels. Although the correlation just missed significance at C1 level (r = .41, p < .05), these results could be interpreted in light of the fact that when the sample size was split into three (i.e. 18, 21, 21), the observed relationship decreased for each group and the correlation lost its significance.

8.3.2.2. Repair fluency

The present study found that total number of repairs were positively correlated in L1 and L2 (r = .44).This lends support to the findings reported in De Jong et al. (2015) and Peltonen (2018), suggesting that L2 repair fluency could be, to some extent, a function of personal speaking styles of individuals. In fact, previous research has demonstrated that this aspect of fluency might better reflect speaker-specific characteristics or individual differences (Baker-Smemoe et al., 2014; Kahng, 2014) since, in several studies, repair behaviour did not demonstrate any change over time (e.g. Di Silvio et al., 2016; Huensch & Tracy-Ventura, 2017), nor did it distinguish between different proficiency levels (e.g.

Baker-Smemoe et al., 2014; Kormos & Denes, 2004). Indeed, in the current study, when correlations were further examined across different proficiency groups, the significant correlation between L1 and L2 repair measures was only maintained at B1 (r = .50) and B2 (r =.55) levels but not at C1. This result also lend support for the findings of Study 1, where total repair in L1 and L2 correlated only at B1 level (r = .60) and those of Tavakoli et al. (2020) which reported that repair measures did not indicate a linear relationship between speech fluency and speaking proficiency levels.

As discussed in Section 2.5.1.2, it is assumed that gains in LP lead to an increased ability to monitor and repair speech, which is likely to result in decreased repair behaviour (Huensch & Tracy-Ventura, 2017; Mitchell et al., 2015). Drawing on this assumption, one could normally expect a stronger relationship between L1 and L2 repair fluency behaviour for learners of higher LP levels. Contrary to such an expectation, however, the results here did not show any correlations for the highest LP level, i.e. C1. As Huensch and Tracy-Ventura (2017) explain, this might be linked to the possibility that while learners improve in their LPs, they simultaneously improve their ability of noticing their mistakes. As a result of this increased awareness of their mistakes, learners tend to reformulate their speech more often and produce more indices of repair (e.g. false starts). This might explain why the correlations failed to show any relationship for repair behaviours in L1 and L2 for the highest LP level.

When interpreted in conjunction with the findings of Study 1, these results overall imply that indices of repair in L2 speech indicate speakers' personal styles, at least to some extent. This finding particularly has important implications for second language practices such as L2 assessment in classrooms or in international language tests, where peoples' use of repair is usually linked with their engagement in the monitoring processes in L2 speech production and the development of their interlanguage system. In such contexts, L2 repair behaviour is often perceived as a sign of disfluency. However, as discussed in Study 1 while repair measures could indicate monitoring processes involved in speech production (e.g. reformulating a structure or correcting errors), they can also reflect, at least partly, speaking styles of individuals. While these findings need to be replicated in other studies, they offer some immediate implications for L2 teaching training programmes as well since L2 teachers' understanding of the relationship between L1 and L2 fluency plays a crucial role in their practices and how they could help learners develop fluency (Tavakoli & Hunter, 2018).

8.3.2.3. Speed fluency

As for articulation rate and speech rate, correlational analysis further indicated significant positive correlations between L1 and L2 fluency behaviour, suggesting that L2 speed fluency reflects some degree of L1 personal speaking styles. In fact, the highest correlation observed in the correlational analysis was for the articulation rate (r = .51). This echoes the findings of De Jong et al. (2015), De Jong and Mora (2019) and Huensch and Tracy-Ventura (2016), all of which examined inverse articulation rate (i.e. mean syllable duration) and found significant correlations between L1 and L2 fluency behaviour. In Huensch and Tracy-Ventura (2016) and Derwing et al. (2009), it was reported that L1 predicted speed fluency from early on, before residency abroad; however, at later stages it predicted speed fluency only for some L1 groups; i.e. Spanish group in Huensch and Tracy-Ventura (2017) and Slavic group in Derwing et al. (2009). While the differences in these studies could be attributed to different degrees of LP development across L1 groups, it is also possible that the closeness of L1 and L2 fluency characteristics, from a typological point of view, might have affected the amount of gains made in different languages in these studies. However, given that the present study focused on only one L1 group, i.e. Turkish native speakers, it seems rather difficult to establish with any certainty whether, and to what extent, the significant correlation for the articulation rate is due to individual speaking styles or cross-linguistic effects. Clearly, further research is needed with more L1 groups to examine cross-linguistic effects on L2 fluency and with different LP levels to understand whether these effects persist across different levels.

Further, a positive significant correlation was also observed between L1 and L2 speech rate, which is a composite measure incorporating pausing and speed fluency behaviours. This finding is in line with the findings of previous research (Bradlow, Kim, & Blasingame, 2017). As typically reported in studies, this result particularly implies that speech rate in L2 occurs in part due to individuals' L1 speaking rate or 'trait' characteristics of speakers (Derwing et al., 2009). In other words, a slow L1 speaker is likely to be slower in their L2 as well. However, interestingly the results for both articulation rate and speech rate are in contrast with the findings of Study 1, which did not indicate any correlations between L1 and L2 for these fluency measures. The discrepancy between Study 1 and 2 could be explained by the amount of language exposure that the participants had had in each of the studies. The participants in Study 1 had not resided abroad for any significant amount of time, whereas the participants in Study 2 came from a study-abroad context and therefore the majority of them had had substantial L2 exposure and experience of living in the L2 speaking community. Indeed, Study 2's findings with respect to speech rate confirms those of Towell et al. (1996) and Derwing et al. (2009), which were also situated in a study-abroad context and reported correlations for speech rate between L1 and L2. Given the overwhelming research evidence suggesting that learners benefit from L2 experience particularly in speed fluency is (e.g. Di Silvio et al., 2016; Du, 2013; Huensch &Tracy-Ventura, 2017; Towell et al.,1996), one could assume that when L2 learners have more opportunities to engage in authentic communication situations, their speech becomes faster, resembling their L1 speech. This hypothesis could offer an explanation for significant correlations between L1 and L2 speech rate for learners in study-abroad contexts such as in the present study.

Related to this, several studies have also demonstrated that speed fluency is one of the most reliable indicators of LP improvement (e.g. Huensch & Tracy-Ventura, 2017; Towell et al., 1996) and therefore one could hypothesize that L1-L2 fluency links would change as LP improves. The participants of Study 1 were overall less proficient (A2, B1 and B2 levels) than Study 2's participants (B1, B2 and C1 levels). Learners at lower LP levels often speak at a slower rate, meaning that they typically have a slower speech rate in their L2 than in their L1. Thus, the significant correlations observed in the current study for speed fluency could also be attributed to higher LP levels of the participants (De Jong & Mora, 2019; Huensch &Tracy-Ventura, 2016). Indeed, the results obtained from the correlational analysis across proficiency groups lend further support for this; it was found that both articulation rate and speech rate in L1 and L2 correlated only at higher-levels of proficiency (i.e. B2 and C1) but not at the lowest level (i.e. B1). De Jong and Mora (2019) argue that speed fluency is not associated with personal styles but rather with L2-specific LP. However, it may also be the case that since processing difficulties encountered in L2 speech production process decreased with increased LP, the influence of other variables such as L1 fluency characteristics might have surfaced. This could offer another explanation for the significant correlations found for higher-level LP levels.

8.3.3. L2 proficiency and L1-L2 fluency relationships

RQ1a set out to explore whether the relationship between L1 and L2 fluency behaviour is mediated by LP. In order to examine this, partial correlations for the significant results obtained through Pearson correlations were carried out, with LP as the controlling variable. It was already discussed in Section 2.5.1 that when learners expand their linguistic knowledge, their automaticity in their speech processing increases, making their speech more similar to their L1s. Therefore, it was hypothesized that at higher levels of LP, one could expect a stronger relationship between L1 and L2 fluency measures. However, the results from the partial correlations indicated that these relationships were maintained across different proficiency levels. This means that overall, LP did not impact on these relationships, except a very small effect on the relationship between L1 and L2 speech rate. These findings confirm the results of Study 1, which did not show significant influence of LP on the strength of the L1-L2 links. However, since these two studies are the only studies which examined whether LP was a mediating factor in L1-L2 fluency relationships (for all aspects), it is impossible to compare the findings with those of any other research. One exception that should be mentioned here is Bradlow et al. (2017), which explored fluency across ten different L1 groups, in terms of speed only. In this study, when the participants' LP was controlled for, the results did not demonstrate any effect of LP on the strength of the links between L1-L2 speech rate across any of the L1 groups. This is overall similar to the current study; although some impact of LP on the correlation between L1 and L2 speech rate was observed (with an increase from .39 to .43), this was very small. Therefore, it could be suggested, at least with some confidence, that although L2 speech rate has widely been reported to be slower than L1 speech, part of this slower rate could be attributed to speakers' speaking styles, regardless of their L2 proficiency.

Previous research has suggested, though not explicitly discussed, that L1-L2 relationships might change as learners' L2 proficiency improves (e.g. Derwing et al., 2009; Huensch &Tracy-Ventura, 2016; Riazantseva, 2001). It was reported in Huensch and Tracy-Ventura (2016) that the relative explanatory power of LP over L2 fluency changed with time; i.e. no influence of LP was observed after learners resided abroad although their LP improved. Similarly, Riazantseva (2001) argued that one's unique pausing patterns in L1 are likely to be transferred to their L2 speech once a certain LP level has been achieved. Such studies are particularly compelling as they clearly suggest that LP influence on L2 fluency is likely to change as LP improves. Therefore, the finding of the current study that the significant correlations persisted regardless of LP seems to be in contrast with the suggestions of previous studies. This finding might again be related to the fact that as L2 linguistic knowledge expands and L2 processing difficulties decrease, the influence of L1 fluency characteristics in L2 speech production process could be observed more.

This line of reasoning could also lead us to interpret LP impact on L1-L2 fluency associations from a different perspective. In structurally similar languages, at earlier stages of L2 proficiency, the relationships between L1 and L2 could be expected to be stronger than at higher-levels, possibly due to L1 transfer effects. Indeed, in Derwing et al. (2009), correlations between L1 and L2 fluency were stronger for L1 Slavic group (which was the closer language to English) than they were for Mandarin group at initial stages; however, the correlations were maintained only for Slavic group at later stages. This led Derwing et al. to conclude that It cross-linguistic effects persist at higher levels for structurally similar languages (i.e. Slavic and Mandarin in their case), and the relationships were maintained due to the similarities between the two. One alternative explanation could be that the degree of the effects of cross-linguistic similarities and LP on L2 fluency might change as LP increases. Huensch and Tracy-Ventura (2016) reported that the explanatory power of these variables over L2 fluency shift with time. Thus, it might be that not only the influence of LP but cross-linguistics effects as well, which are arguably stronger at lower LP levels (i.e. due to transfer effects), become less observable at higher LP levels. Consequently, it might again become possible to observe the influence of other factors, such as L1 fluency. However, as Huesnch and Tracy-Ventura (2016) also explain, all these factors (LP, L1 and cross-linguistic effects) contribute to L2 fluency- and subsequently on the L1-L2 relationships- to different extents. Therefore, these factors should be considered in conjunction. It is also important to bear in mind that different L1 groups and LP levels are also likely to result in different findings. As such, caution is needed when interpreting the findings here.

8.3.4. WMC and L1-L2 fluency relationships

RQ1b asked whether the relationship between L1 and L2 fluency behaviours is mediated by individual variations in WMC. Based on the evidence from previous research, it was hypothesized that a higher WMC would support L2 speech production especially at lower-LP levels, and thereby L1-L2 fluency relationships would be mediated by variations in WMC. To answer the RQ, partial correlations controlling for WMC were carried out for the significant correlations obtained through simple correlational analysis. The results overall indicated that the relationships between L1 and L2 for all fluency measures were maintained, with no influence of individuals' WMC. Some impact of WMC was observed (i.e. a small decrease from .51 to 46) only for the relationship between L1 and L2 articulation rate. This effect was very small though. In fact, given that WMC effects have been implicated particularly for speed fluency in the relevant literature (e.g. Gilabert & Munoz, 2010; Mota, 2003), one could have expected to find even a bigger impact of WMC for this measure. Still, this small decrease in the strength of the correlation between L1 and L2 speed fluency may imply that WMC mediates the L1-L2 fluency links, to a small degree, with respect to speed aspect only. In this regard, it seems plausible to suggest that those participants with a higher WMC better managed their attentional resources and produced faster speech. Related to this, it is also possible to assume that since individuals with a higher WMC arguably had a higher degree of automaticity (Segalowitz, 2010), their increased ability in lexical access resulted in faster speech rate (Mota, 2003). However, one should also bear in mind that the WM was measured in L2 rather than L1 in the present study. Although this was motivated by previous research which reported that WMC scores in L1 and L2 correlated significantly with each other (e.g. Awwad & Tavakoli, 2019), it might be that the participants in previous research were at similar level of proficiency whereas in the present study they had different levels. Therefore, LP might have confounded the results for WMC tests as a result of this limitation. In fact, LP could have been a mediating factor in WMC-fluency links as well given that WMC contributes to L2 fluency performance differently across different LP levels as was implicated in earlier studies (Georgiadou & Roehr-Brackin, 2017). The results, therefore, should be interpreted with a consideration of a possible interaction effect of LP and WMC on L1-L2 fluency relationships. Since no previous study, to the best of my knowledge, has examined whether individual variations in WMC mediate L1-L2 fluency relationship, it is not possible to compare these results with those of others. However, based on the findings here, one could suggest with some confidence that the relationships between L1 and L2 fluency behaviours persist regardless of the effects of individual variations in WMC.

8.3.5. LoR and L1-L2 fluency relationships

Another RQ was formed to examine whether the relationship between L1 and L2 fluency behaviour is mediated by learners' LoR. In this study, LoR was considered as a proxy for overall L2 linguistic experience in the TL community, i.e. UK. In order to answer the RQ, further partial correlations were carried out for the significant correlations obtained through Pearson correlations, this time controlling for variations in LoR. Drawing on the existing research evidence suggesting that L2 speech fluency benefits most from L2 experience abroad (e.g. Du, 2013; Huensch & Tracy-Ventura, 2017; Kim et al., 2015; Mora & Valls-Ferrer, 2012), it was assumed that the amount of time spent abroad would have an impact on the strength of L1-L2 fluency relationships. However, the results from the partial correlations overall did not indicate any effect of LoR on these relationships. Only for the correlation between L1 and L2speech rate, there was a small increase (from .39 to .45). Based on the research evidence about LoR particularly in favour of speed fluency (Di Silvio et al., 2016; Du, 2013; Huensch &Tracy-Ventura, 2017; Towell et al., 1996), this finding does not seem surprising. Although the influence of LoR was small, it implies that when this effect was partialed out from the correlation, the relationship between L1 and L2 speech rate becomes even stronger. This might be explained in the light of the fact that when learners engage more in authentic linguistic activities in a TL community with native and non-native speakers on a daily basis, they have more opportunities to become exposed to language input. As a result, a tentative explanation might be that learners tend to change their speaking style with respect to speech rate to suit more to the norms of TL, and this could be reflected in their L1 personal speaking styles as well. These results also suggest that the relationship between L1 and L2 speech rate, which combines breakdown and speed fluency aspects, is more likely to be affected by L2 linguistic experience during study abroad than other fluency relationships.

There may be other possible explanations for the surprising finding that LoR did overall not have an impact on the strength of the L1-L2 correlations. As discussed earlier in Section 2.5.3, it does not necessarily mean that all learners spend their time abroad interacting in the TL (Flege & Liu, 2001). To what extent fluency development benefits from residency abroad is largely dependent on whether individuals opt to use their L1 or L2 mostly as their main language for communication or on their voluntary exposure to the language, e.g. by voluntarily and purposefully engaging in communitive activities (Derwing, Munro, & Thomson, 2008). Therefore, it may be that as a result of how individuals spend their time abroad, they are likely to improve in their fluency to different extents. Hence, such variations in fluency gains might have influenced the results here.

Finally, a lack of LoR effect on the L1-L2 fluency associations might also be linked to an interaction effect of LP and LoR. Although LP can improve along with increased LoR (e.g. Hernandez, 2010), the longer amount of time spent abroad does not necessarily lead to higher LP because, similar to fluency gains, individuals are very likely to differ in terms of gains in overall proficiency. Given that the participants in Study 2 came from higher LP levels, each of these factors, i.e. LP and LoR, might have influenced the results to different degrees. While this point is worthwhile to consider when interpreting the results here, an investigation of the possible interaction effect was beyond the scope of the current study.

8.4. Lexical complexity in L1 and L2

Although an overriding focus of Study 2 was investigating L1 and L2 fluency links, RQ2 and RQ3 of the study were formed to explore lexical complexity behaviours in L1 and L2 speech. In this section, I discuss the findings related to these RQs. I start with observations for the differences between L1 and L2 lexical complexity and then move on to a discussion of the findings from Pearson correlational analysis. Similar to what was done in the previous section, I attempt to compare the results with those of earlier studies and offer some explanations for the findings from the literature.

8.4.1.Differences in L1 and L2 lexical complexity

The descriptive statistics highlighted large differences for LD measures, i.e. TTR values and D scores, between L1 and L2. The participants had higher LD scores for both measures in their L1 than their L2, with TTR values of L1 M = .56 and L2 M = .44, and D scores of L1 M = .49 and L2 M = 23. The results of the t-tests showed that the differences between L1 and L2 scores for both LD measures were statistically significant. This suggests that the participants produced more lexically diverse speech in their L1 Turkish than L2 English. The figures are in line with the findings of previous research in that the speakers exhibit different lexical patterns across languages, most often in favour of L1 and/or the most proficient language (e.g. Daller et al., 2003; De Clerq, 2015; Dewaele & Pavlenko, 2003; Johansson, 2009). Dewaele and Pavlenko (2003) reported that Russian speakers of English as EFL and ESL had overall lower LD scores than monolingual American speakers. Further, in Daller et al. (2003), the Turkish-German bilinguals had higher LD scores in their dominant language; i.e. the German-dominant group had higher scores in German while Turkish-dominant group had higher scores in Turkish. The participants in Daller et al. (2003) were also shown to be more proficient in their dominant language, lending further support for the findings of the current study.

The findings of the current study also seem to confirm that LD, as one aspect of speech performance, is indicative of proficiency in one given language (Daller et al., 2003; Treffers-Daller, 2013). Earlier in Section 2.2.2., it was discussed that L2 mental lexicon is not yet as extensive and well-organized as L1 lexicon (Skehan, 2015), and it develops with increased linguistic knowledge. As such, linguistic resources (morphological and syntactic) available in the lexicon of a given language impact how lexical resources are utilised (Strömqvist et al., 2002); therefore, the lower LD scores in this study might be explained by the relatively small and underdeveloped L2 lexicon of the participants here, which might have posed additional difficulties for them in their L2 speech, resulting in lower LD scores.

8.4.2. Correlations between L1 and L2 lexical complexity

RQ3 was formed to explore whether there was a relationship between L1 and L2 lexical complexity behaviours, which were represented through LD scores. The goal originated from the research which have suggested that lexical complexity performance is likely to be affected by several factors including cross-linguistic effects (e.g. Daller et al., 2003; Dewaele & Pavlenko, 2003) or L1 stylistic variations (Pallotti, 2015). It was hypothesized that personal preferences or speech styles in L1 could, at least partly, be reflected in L2 lexical complexity performance. In order to answer the RQ, Pearson correlations were carried out between LD scores, i.e. TTR values and D scores, in L1 and L2. Following from Treffers-Daller et al. (2018), TTR values were used as a traditional method of examining LD whereas D measure was employed as one of the more recent (corrected) LD measures. The results indicated a significant positive correlation between L1 and L2 LD for TTR values only (r = .41) while they did not show any correlations for D scores. The different results for TTR values and D scores may be explained by the differences in the computations of these scores; while the calculation of TTR values were based on simple counts of types and tokens in the participants' speech samples, D measure, though essentially based on the TTR values, provided a score of LD based on a series of different computations (i.e. taking the text lengths of 35-50 tokens and computing a random sampling of TTR curves). Previous research has suggested that the traditional method (i.e. simple counting of types and tokens) could better detect within and between group differences than corrected measures as D measure (e.g. Treffers-Daller et al., 2018). Hence, it could be argued that the TTR values provided more reliable results in estimating the LD in participants' oral performance.

To the best of my knowledge, no previous research has explicitly examined the relationship between L1 and L2 lexical complexity. Still, the results of the present study lend support for the suggestions made in the relevant literature. The significant correlation between L1 and L2 TTR values seem to be in the line with Dewaele and Pavlenko (2003) and Pallotti (2015), which suggested that personal stylistic variations may be at play in lexical complexity performance. As discussed in Section 5.6, it is possible to observe stylistic lexical variations in people's L1 speech, and such styles might partly be reflected in their L2 speech as well. The moderate positive correlation found for L1 and L2 TTR values implies that the participants who used more diverse vocabulary in their L1 speech also produced more diverse vocabulary in their L2 speech. De Jong and Mora (2019) argued that individual variations in the use of non-verbal skills (e.g. the non-verbal intelligence called on in conceptualization stage) involved in speech production process could account for some of individual differences in oral performance. Given the inter-relatedness of the speech production processes (e.g. lexical retrieval and articulation of the overt speech) with each other, it could arguably be suggested that such non-verbal individual skills might have been involved in the formulation stage as well. In other words, those participants who had better individual skills, which are language-independent, might have had faster retrieval skills. Arguably, this might also have helped them access a wider range of different words from the lexicon in both their L1 and L2.

Additionally, the findings could be interpreted in light of the fact that some individuals might not use a wide range of different words; however, this does not necessarily mean that their lexicon is not rich enough or readily accessible but might simply indicate that these people do not prefer to put their knowledge into use (Laufer, 2003). Such preferences might have been reflected in the participants' speech here; i.e. it might be that those who used a wider range of words in their L1 speech also used more diverse vocabulary in their L2s, partly due to their styles. It is important, however, to keep in mind that these results may be limited by the fact that the current study examined only LD aspect of lexical complexity performance, and only Turkish as L1. Clearly, future research needs to further investigate possible relationships between L1 and L2 for other aspects (i.e. lexical sophistication and lexical density) with bigger samples sizes and with different L1 groups. Another important point that should be brought to the attention is that the lexical analysis here did not consider the formulaic language use. As mentioned in Section 2.4.2 , this is a limitation of most studies in this area and the current study is no exception. Given that the use of formulaic language in speech has been related with advanced LP (Tavakoli, 2011; Tavakoli & Uchihara, 2019), this issue surely deserves further consideration in future lexical studies.

8.4.3. Correlations for fluency and lexical complexity measures

The aim of the last RQ in Study 2 was to explore whether fluency and lexical complexity behaviours are related to each other in L1 and L2 separately. In order to address the RQ, further sets of Pearson correlations were carried out between fluency and LD measures in each language separately. As was discussed earlier, from a psycholinguistic point of view, lexical retrieval problems faced by speakers in the formulation stage of the speech production process or conscious attempts made at using more diverse vocabulary in speech are likely to result in the disruption of the smoothness of the speech and disfluencies in the overt speech, which is especially the case for L2 speakers. Based on this, it was assumed that lexical complexity and fluency aspects of oral performance would be related to each other, at least partly.

The results of the correlational analysis indicated relationships only between speech rate and TTR values in both Turkish (.39) and English (.39) (except articulation rate which also correlated with TTR values in Turkish). The results did not show any relationships between any of the fluency measures and D scores. When the relationship between speech rate and TTR values were further investigated across LP levels, it was observed that the correlations surfaced only for learners at higher L2 levels (i.e. C1) in both languages. Previous studies have reported that speakers are more likely to pause before low-frequency words when compared to high-frequency words (e.g. Bosker et al., 2014; De Jong, 2016), suggesting a link between fluency and lexical complexity behaviour. The findings of the current study broadly seem to support these studies in that they suggest a link between the two aspects; however, since the analysis of lexical complexity in the present study was limited to LD aspect only, it is difficult to claim whether the use of low or high frequency words could be associated with pausing behaviour. Clearly, an in-depth analysis of low and high frequency words in L1 and L2 speech data is needed to understand this. This is a point for future studies to investigate.

Additionally, it was argued that if there is a link between these two aspects of the oral performance (i.e. fluency and lexical complexity), there is little evidence to suggest that this relationship is linear or in one direction. In other words, it is not clear whether more pausing and slower speech leads to decreased or increased LD in speech (although one argument could be that it should result in the use of richer vocabulary, e.g. as a consequence of more carefully-generated speech). However, the results of the present study imply that LD in speech increases with speech rate, regardless of the language spoken, i.e. L1 or L2. In general, these findings seem to support those of De Jong and Mora (2019) who reported that bigger vocabulary size is linked with faster articulation rate. These findings could be related to individual speaking styles, which have been suggested to be influential in both L1 and L2 pausing (e.g. De Jong, 2016; De Jong et al., 2015; Peltonen, 2018) and lexical complexity behaviours (e.g. Pallotti, 2015). The positive moderate correlation found between L1 speech rate and L1 LD implies that speakers who have a faster speech are also those who produce a wider range of different words in their speech.

However, it was interesting to observe that the correlations between speech rate and TTR scores surfaced only at higher LP level (i.e. C1), but not at lower levels, in both of the languages. This might be explained by the nature of L1 and L2 mental lexicons and the automaticity in lexical access; i.e. mostly automatic lexical access to a well-developed lexicon in L1 speech and mostly less automatic lexical access to a smaller and less-developed lexicon in L2 speech (Kormos, 2006). It is suggested that as linguistic knowledge improves, L2 speakers become faster in their speech and use more diverse vocabulary, i.e. they improve in fluency and LD. Given that L2 learners improve in both of these aspects, this could arguably lead to a higher association between speech rate and LD at higher LP levels. However, an interesting result emerging from the analysis was that a similar pattern was evident in Turkish language; i.e. the participants who were more proficient in L2 (i.e. at C1 level) spoke faster and used more diverse vocabulary in their L1 as well. When these results are interpreted in conjunction, one could suggest that speech rate and LD were found to be linked to each other because both of these aspects are likely to be a function of personal speaking styles, at least to some extent.

8.5. Summary of the chapter

In this chapter, I have discussed the findings of Study 2 as they relate to the two aspects of second language oral performance: fluency (RQ1 and RQ1a, RQ1b and RQ1c) and lexical complexity (RQ2 and RQ3). In the first section, I explored the relationship between L1 and L2 fluency behaviour and whether this relationship is mediated by variations in individual learner variables; i.e. LP, WMC and LoR. I linked the findings emerging from simple and partial correlations to the relevant literature in this area. I suggested that L1 and L2 fluency behaviours are linked to each other, at least to some extent, for all fluency aspects; i.e., breakdown, repair and speech fluency. Based on the strengths of the correlations between fluency measures in both languages, I argued that particularly repair and speed fluency in L2 might reflect L1 speaking styles. I also discussed that these relationships overall appear to be independent of variations in LP, WMC and LoR and noted that the differences in these variables affected L1-L2 fluency links to a very small extent only when speed fluency was concerned. I explained that since these factors had been shown to have a particular impact on L2 speed fluency, it was likely that they could influence the associations between L1-L2 speed fluency as well although the observed impact of each was very small.

The second section explored the relationship between L1 and L2 lexical complexity behaviour represented through LD, and examined whether the two aspects of oral performance, i.e. lexical complexity and fluency, were related to each other in L1 and L2 separately. Here I explained that when measured through TTR values, LD in L1 and L2 were linked with each other to a small extent. I also reported that these aspects, i.e., fluency and LD, correlated with each other to a limited extent in each of the languages, i.e., correlations were observed only for speed fluency and TTR values and only at C1 level. I provided possible explanations for this, suggesting that while these correlations could be attributed to the structure of the mental lexicon and the degree of automaticity in each language, it is also likely that personal speaking styles might be at stake in terms of speed fluency and lexical complexity behaviours.

CHAPTER 9. CONCLUSION

9.1. Introduction

In this chapter, I provide the main findings from Study 1 and Study 2. I highlight the significance of these findings for SLA and fluency research and the contributions this thesis has made at both theoretical and methodological levels. Then, the implications of the present research will be outlined in terms of theoretical, methodological and pedagogic domains. To conclude, I mention some of the limitations of the present research, and suggest potential directions for future research in the area.

9.2. Summaries and conclusions

Several studies have suggested that in order to draw a better picture of L2 fluency behaviour, sources of variability in the performance, which are not related to L2 such as L1 personal styles, should be identified and taken out from the speech (e.g. De Jong et al., 2015; Segalowitz, 2010). It is often the case that L1 styles, though non-L2 specific, could influence people's judgements on speakers' L2 fluency. While there has recently been a growing interest in understanding how L1 fluency could impact L2 fluency behaviour (e.g. De Jong et al., 2015; Huensch & Tracy-Ventura, 2016; Peltonen, 2018), research in this area has been very limited. Equally important, this limited body of research has focused mostly on the examination of structurally similar language pairs (e.g. English-Spanish or English-French) and/or worked with only one or two L2 proficiency groups (mostly advanced learners). Study 1 and 2 have been the first to explore the relationship between L1 and L2 fluency behaviours for Turkish learners of English. Although the prime goal of both studies was to examine the possible fluency links between L1 Turkish and L2 English, they differed in several aspects. While Study 1 was situated in an EFL context and explored the fluency behaviours of lower-level learners (i.e. A2, B1 and B2 at CEFR), the participants in Study 2 were on study-abroad degrees in the UK and had higher LP levels (i.e. B1, B2 and C1 at CEFR). Another difference was related to the factors explored in each which could potentially mediate L1-L2 fluency relationships. Study 1 explored the mediating impact of L2 proficiency (described with three levels) and task structure (i.e. tight or loose) on the strength of L1-L2 fluency relationships. Study 2, on the other hand, explored whether L2 proficiency (described with three levels), WMC and LoR (both used as continuous variables) mediated L1-L2 links. A final difference was that Study 1 was exploratory with regard to measures of fluency, using a total of 17 fluency measures, whereas Study 2 built upon the findings of Study 1 and employed a reduced number of fluency measures (five in total). In both studies, participants LP levels were measured through the use of two standardized tests, i.e. OPT and EIT; with the former being assumed to test declarative knowledge and the latter the procedural knowledge.

An additional goal of Study 2 was to explore the links between L1 and L2 lexical complexity as well as between fluency and lexical complexity in each language. The motivation for including the examination of the lexical complexity in Study 2 came from that 1) fluency and lexical complexity were shown to be two of the most reliable indicators of L2 proficiency (e.g. Iwashita et al., 2008; Revesz et al., 2016), 2) the study was built on the assumption that fluency and complexity aspects were inter-linked to each other; e.g. lexical access or retrieval problems encountered in the formulation stage of speech production process are likely to manifest themselves as disfluencies in the overt speech, and 3) research has suggested that lexical aspect of the oral performance is also likely to be affected by several factors, including cross-linguistic effects (e.g. Daller et al., 2003; De Clerq, 2015; Dewaele & Pavlenko, 2003) or L1 stylistic variations (Pallotti, 2015). In the following sections, I will summarise the main findings from each of the two studies carried out for this thesis.

9.2.1.Conclusions from Study 1

The investigation of the relationship between L1 and L2 fluency behaviours for learners at lower levels in an EFL context has clearly demonstrated that L2 fluency is, at least to some extent, a function of L1 personal speaking styles. Breakdown (i.e. mid-clause pausing behaviour) and repair fluency (i.e. number of reformulations and total repair) were found to be linked in L1 and L2 for lower-level learners, suggesting that L2 fluency could partly reflect L1 styles in these aspects of fluency. These findings overall lent mixed support to those of previous studies (De Jong et al., 2015; De Jong & Mora, 2019; Derwing et al., 2009; Huensch & Tracy-Ventura, 2016; Peltonen, 2018) although there was not unity in the measures used across these studies. The results for speed fluency and the composite measures, on the other hand, did not show any links between the two languages and this was in contrast to the results reported in previous studies. Another important finding which has emerged from Study 1 was that L1 and L2 fluency relationships were not mediated by the structure of the tasks (i.e. loose or tight) or individual variations in L2 proficiency. These findings have overall indicated that L1-L2 fluency associations persisted regardless of the influence of these factors.

With respect to the predictive power of L1 fluency and L2 proficiency over L2 fluency, interesting findings have emerged. While L1 fluency predicted L2 breakdown and repair behaviours to a certain degree, L2 proficiency scores, i.e. the EIT and the OPT scores, predicted different aspects of fluency. The OPT scores predicted, to some extent, L2 breakdown and repair aspects, i.e. frequency of mid-clause pausing behaviour (silent) and reformulations. The EIT scores, on the other hand, predicted mostly speed-related fluency (i.e. all composite measures, combining speed and other aspects of fluency) and partly repair and breakdown fluency (i.e. total repair and frequency of mid-clause silent pausing). The findings have suggested that in speech production process declarative knowledge may encourage mid-clause pausing behaviour as L2 learners might be using pauses as opportunities to access to their declarative knowledge stores whereas procedural knowledge seems to be a good predictor of an uninterrupted and fast speech. These results have also highlighted the importance of choice of test in studies when assessing L2 proficiency of learners since different tests used seem to tap into different aspects of knowledge, i.e. declarative and procedural linguistic knowledge.

9.2.2.Conclusions from Study 2

The results from Study 2 have revealed similar findings to those from Study; L1 and L2 fluency behaviours were related for learners at higher LP levels as well. Unlike Study 1, Study 2 found relationships for all aspects of fluency; i.e. breakdown (i.e. number mid-clause silent pauses), repair, speed and composite measures. This suggests that not only at lower levels of LP as the results from Study 1 indicated, but at higher levels as well, L2 fluency mirrors, to some extent, L1 personal speaking styles. These results have overall confirmed the findings of Study 1 with regard to breakdown and repair aspects, and those of other studies of this kind (e.g. De Jong et al., 2015; Huensch & Tracy-Ventura, 2016; Peltonen, 2018). When the influence of each of the individual learner variables (i.e. L2 proficiency, WMC and LoR) was controlled for, L1-L2 relationships were maintained. Taken together, these findings have clearly revealed that L2 fluency behaviour is, at least to some extent, related with L1 fluency, and these relationships were overall not influenced by variations in individual learner variables; namely, L2 proficiency, WMC and LoR.

As for the relationships between L1 and L2 lexical complexity behaviours (represented through LD only), the study has found different results for the performance between the two languages measured by TTR values and D scores. L1 and L2 lexical performances were related to each other only to a small extent when the performance was measured by the TTR values. Although no previous research has explicitly examined the links between L1 and L2 lexical complexity behaviours, the results lend support for the suggestions made in the literature (e.g. Dewaele & Pavlenko, 2003; Pallotti, 2015) in that L1 personal stylistic variations may be at play in L2 lexical performance as well. The findings have also suggested that the traditional method, i.e. TTR values (simple counting of types and tokens) could perhaps better detect within and between groups differences and should therefore be considered as a more reliable method in estimating the LD aspect of oral performance than the corrected measure of D scores.

Finally, Study 2 found that fluency and lexical complexity behaviours in each language separately are only to a limited extent related to each other. The results showed links only between TTR values and speech rate in both English and Turkish (with the exception of Turkish articulation rate which was also found to be related with Turkish TTR values).

Overall, the findings have suggested that speakers who have a faster speech rate are also likely to produce a wider range of different words in their speech.

9.3. Contributions of this research

In this section, I would like to reflect on what has been achieved in this thesis. There is no doubt that there remains much work to be done in SLA and fluency research. However, the findings from the two research studies presented here have made several valuable contributions to the current literature at both theoretical and methodological levels.

To start with, the current research has been the first to explore whether fluency behaviours in L1 Turkish and L2 English are related. Research studies examining the role of L1 fluency in explaining L2 fluency behaviour have been very scarce. The current research has addressed this gap in several aspects. First, it has explored the relationship between L1 and L2 fluency behaviours for two typologically distant languages, i.e. L1 Turkish and L2 English, unlike most other previous research which focused on examining structurally similar language pairs such as L1 English- L2 Spanish in De Jong and Mora (2019) or L1 English-L2 French in Huensch and Tracy-Ventura (2016). This is an important contribution of this research to the area because the influence of L1 fluency on L2 fluency could be reflected not only through L1 personal speaking styles, but cross-linguistic similarities and differences between the languages in question as well. This means that in typologically or structurally similar languages, learners might find it easier to improve their L2 fluency than those who have very distant L1s and L2s. From this perspective, although this thesis has not focused on examining cross-linguistic effects yet examined two typologically distant languages (Turkish and English), the emerged findings could shed some light on the role of L1 background, be of personal speaking styles or cross-linguistic influences, in explaining L2 fluency.

Further, no prior research to date has examined the mediating role of L2 proficiency in the strengths of L1-L2 fluency relationships. Study 1 and Study 2 have been the first two studies, to the best of my knowledge, to address and explore this. Perhaps more importantly, in both studies L2 proficiency has been investigated from a broader perspective than in previous studies. This has been achieved in two ways. First, L2 proficiency has been described with three groups in each study (i.e. A2, B1 and B2 in Study 1 and B1, B2 and C1 in Study 2) unlike previous studies which worked with L2 learners at one or two proficiency levels. Given the crucial role of LP in fluency development, knowing how learners behave across different LP levels including low-levels was even more necessary. As such, Study 1 was aimed for learners at lower levels while Study 2 for learners at higher levels of proficiency. Second, LP assessment has often been unsystematic in previous research with some using a vocabulary test or others using no test at all but basing their research on school levels. A single test such as grammar or vocabulary test could provide interesting insights into learners' proficiency; however, it falls short of indicating much about their speaking skills. Similarly, a speaking test, such as EIT, would be limited in drawing a picture of learners' overall linguistic knowledge. An important methodological contribution of the current research is that L2 proficiency has been assessed systematically, using two standardized proficiency tests of OPT and EIT. The two measures were used to capture individual variations amongst the participants more comprehensively by tapping into different underlying constructs of L2 proficiency; namely, the OPT to measure L2 declarative linguistic knowledge (mostly) and the EIT to measure L2 procedural linguistic knowledge (mostly). This more systematic approach to assessing LP has contributed to gaining a more complete and reliable profile of the learners' linguistic development, and thus shed more light on the role of LP in L1-L2 fluency associations. It has also helped us see which aspects of fluency were linked to each type of linguistic knowledge. It was found that declarative knowledge was likely to be related with mid-clause pausing behaviour while procedural knowledge encouraged speed fluency. These findings will be of interest especially to SLA researchers as they clearly indicate that the choice of test to measure LP is likely to have an impact on the results.

In addition to the examination of the role of L2 proficiency, the current research has also been the first attempt, to the best of my knowledge, to look into the influence of other factors, both learner internal and external, which could potentially mediate L1-L2 fluency relationships. Study 1 has been the first to explore whether task structure as an external factor had any impact on the strength of L1-L2 links while Study 2 has examined the influence of other individual learner variables, i.e. WMC and LoR. Each of these factors has been shown to influence L2 fluency in different ways as was already discussed in Section 2.5; therefore, they were predicted to influence L1-L2 associations by either debilitating or facilitating L2 fluency performance. This research has taken a step towards exploring the mediating impacts of these variables on the relationships between L1 and L2 fluency behaviours, but it has found none.

Compared to previous studies of this kind which have centred on L2 learners who are either on study-abroad courses or immersed in the TL community, the present research has investigated performance in both EFL and study-abroad contexts. While study-abroad studies are valuable as they provide interesting insights into understanding L2 fluency performance or development, it is equally important to conduct studies in EFL contexts, where learners do not have many opportunities to have extensive exposure to the TL or to use the language for authentic purposes. By situating Study 1 in an EFL context and Study 2 in a study-abroad context, this thesis has not only contributed to (lack of) the existing research, but has provided a unique opportunity to compare the results for learners who studied L2 English in different contexts, yet who came from a similar educational background in their country and spoke Turkish as L1.

Finally, for the first time in the literature, Study 2 has explored the link between L1 and L2 lexical complexity behaviours. My reading on the relevant literature has suggested that lexis aspect of performance is also sensitive to cross-linguistic effects (e.g. typological differences across languages) and individual differences (e.g. such as L1 stylistic variations). How much lexically rich and linguistically complex information is given in speech seems to be dependent on various factors (e.g. familiarity with the topic, having specialized knowledge, etc). Study 2 aimed to explore one of these, i.e. the role of L1 in L2 lexical complexity, by examining the relationship between lexical performances in both languages. Although this was done in a rather limited way for reasons which I will explain in Section 9.5 below, Study 2 appears to be the first study examining L1 Turkish and L2 English in this regard, and clearly contributes towards enhancing our knowledge of L1 influence on L2 oral lexical performance.

9.4. Implications of this research

Drawing on theories related to speech production processes, fluency research has primarily focused on the investigation of temporal features in speech (such as pauses and hesitations) as a way to uncover the psycholinguistic processes underlying L2 fluency performance and development (e.g. gaining automaticity). In the majority of studies, increases in L2 fluency were attributed to increases in linguistic knowledge or degree of automaticity in that knowledge. This has led L2 researchers to draw conclusions about speakers' performance based on fluency indices in their studies; speakers with fewer pauses or corrections were perceived as more fluent, and maybe more proficient. However, recent research has raised interesting questions about how much of L2 fluency could be attributed to L2-specific performance or to what extent L2 fluency is a characteristics specific to an individual or to a language. The current research has been an attempt to shed some light on these questions and to explore the role of L1 fluency in explaining L2 fluency behaviour. A number of theoretical, methodological and pedagogical implications have emerged from the findings, which are discussed in the following sections.

9.4.1. Theoretical implications

When Levelt's speech production model in Section 2.2.1 was discussed, it was argued that this research was mainly built on the assumption that disfluencies observed in the overt speech are likely to be the result of processing difficulties (e.g. lexical retrieval problems, grammatical encoding, etc.) encountered in the formulation stage. I explained that this is especially the case in L2 speech since learners' linguistic knowledge is not extensive or yet automatic. The findings from both Study 1 and Study 2 imply that fluency indices such as frequent pauses, hesitations or slower speech could not be purely an L2 specific phenomena which occurs due to processing challenges faced in speech production process or due to the underdeveloped nature of learner's interlanguage as seen by most L2 practitioners (be them of researchers, testers or teachers). Rather, such disfluencies could be a function of L1 personal speaking styles, at least to some extent. In other words, a person who produces a lot of filled pauses, for instance, in their L1 speech is likely to exhibit a similar amount of pause in their L2 speech as well. The current findings support the suggestions made in previous research (De Jong et al., 2015; Segalowitz, 2010), and provide us a better understating of L2 fluency performance. Perhaps more importantly, they could make valuable contributions to the development of a better-informed L2 speech production model. Clearly, any L2 speech production model should accommodate the effects of L1 fluency behaviour when characterising fluent speech.

The findings have also revealed that L1 styles are carried over to L2 speech mostly in terms of breakdown and repair aspects. As for breakdown fluency, pausing at clause boundaries, no matter which language is spoken (i.e. L1 or L2), is related to conceptualization issues; i.e. to make a conceptual plan or what to say next. Mid-clause pausing on the other hand is suggested to be a key characteristic of L2 speakers. The findings from both studies here have suggested that frequency of mid-clause pausing could to some extent mirror L1 styles whereas duration of mid-clause pauses in L1 and L2 speech seem to be independent from each other. The latter finding contrasted with the previous research which suggested that pause length could be linked with L1 personal styles but not with L2 proficiency. The current findings, however, imply that pause length could be a characteristic specific to a language rather than to an individual.

Another interesting implication about mid-clause pausing concerns the character of pauses (i.e. filled or silent). It was observed that at lower-levels of LP, the relationships between L1 and L2 mid-clause pausing surfaced in the form of silent pauses whereas at higher-levels they emerged in the form of filled pauses. This suggests filled and silent pauses across different levels might be indicative of different processes taking place in L2 speech production process. It is likely that learners use silent pauses as opportunities to access to their declarative knowledge especially at initial stages of L2. In fact, the results of the multiple regressions seemed to support this; i.e. the scores from OPT, a measure which was used to test declarative knowledge, predicted the frequency of mid-clause silent pauses. On the other hand, the more frequent use of mid-clause filled pauses at higher levels of LP could imply that L2 speakers tend to use filled pauses as communicative strategies, i.e. to seem more fluent by filling silence with non-lexical fillers (e.g. umm, erm) as they expand their linguistic knowledge.

Along with increased knowledge, the relationship between L1 and L2 repair behaviours appears to change as well. It was observed in both Study 1 and Study 2 L1-L2 repair links surfaced only at B1 level but not at higher levels. This consolidates the suggestion of previous research in that repair behaviour does not change in a progressive pattern across LP levels, and therefore it may not be indicative of L2 proficiency. As was mentioned already, a similar pattern was observed for the number of mid-clause filled pauses as well. One implication is the possibility that as learners improve in their linguistic knowledge, they simultaneously improve in their ability for noticing mistakes in speech. The more frequent use of both filled pauses and repair instances by B1-level learners might be serving a similar function in speech; i.e. to compensate for the delay which stems from processing difficulties in the formulation stage.

Finally, with regards to speed fluency, the evidence from this research suggests that L1 and L2 relationships are likely to surface at higher levels, because higher-level learners face less challenges in speech production due to increased linguistic knowledge and degree of automaticity and their L2 speech becomes similar to their L1. This has also been implicated in the finding that speed fluency links between L1 and L2 was observed for learners in the study-abroad context only, but not in an EFL context. Study-abroad experience, which arguably provides L2 learners with more opportunities to expose themselves to the language, has been shown to be beneficial especially for improving speed fluency. Indeed, one could observe that participants in most fluency studies which found links between L1 and L2 speech rate, were either on study-abroad courses or immersed in the TL community. In an EFL context, however, this research did not find any associations between L1 and L2 speed fluency. This highlights the importance of carrying out studies in different contexts, including EFL contexts as well, given that fluency performance develops in different ways and lead to different findings.

9.4.2.Methodological implications

The current findings have a number of methodological implications which may be of value to researchers in the fields of SLA and fluency. The first of these relates to the way pauses are identified. It is common to find studies on L1-L2 fluency relationships which identified pauses at AS-unit boundaries or measured pauses only in terms of total frequency or duration. Yet, making a finer distinction of pauses, i.e. according to mid and end clause positions as in the current research, could provide a deeper insight into pausing behaviour in L1 and L2 speech (De Jong, 2016; Skehan & Foster, 2007; Tavakoli, 2011). As discussed in the previous chapter, the methodological difference in pause identification has made it difficult, even impossible at times, to compare the findings for number and length (silent) of pauses at clause boundaries with those reported in previous studies. I also argued that some of the discrepancies in the results across studies might have been caused by this. The lack of a unified and unanimously agreed approach in employing fluency measures in different studies is both a disadvantage and a challenge L2 research in this area faces; yet, it is hoped that the current findings shed some light on the usefulness of employing measures which could help make a more detailed analysis of speech fluency.

The present research identified silent pauses and pause durations (filled and silent) using computer technology, i.e. specialist PRAAT software (Boersma & Weenink, 2013). As was described in Section 6.11.4, silences within speech were automatically detected with 'Annotate to Textgrid silences' command and tiers were created to note down the filled pauses. Surely, such automatic fluency analysis makes it feasible to deal with a large amount of speech data more objectively and precisely. Yet, it is relatively uncommon to find studies which provide detailed descriptions of how PRAAT was used in their research, the challenges faced or the potential issues with using this software. I would like to note that while the software could be very useful in analysing temporal features of speech such as pause frequency or pause length, as was indicated in other studies (e.g. Cucchiarini et al., 2000, 2002; De Jong & Perfetti, 2011; Hunter, 2017), it requires a very detailed manual inspection of the text grids by the researcher. Although in monologic tasks, such as narrative tasks used in the present study, speech data is expected to be more straightforward and clearer, background noise (e.g. page turns) or irregularities in recording are still likely to feature (Hilton, 2014). It is important that researchers should not entirely rely on silences generated by the tool itself without a closer inspection since the boundaries of silent pauses are often marked very broadly on the spectrogram. The precise locations of pauses need to be carefully checked for all pauses. It is possible that a whole segment which is indicated as speech (sounding) on the spectrogram could actually be a combination of a filled pause and an utterance; for example, when the ending of the filled pause 'erm' combines with the beginning of a similar sound 'made' (erm made it). Similarly, PRAAT may indicate the length of a speech segment with a plosive sound (e.g. voiced p, t or k, or voiceless, b, d or g) at the end as way longer than it is in reality. This is likely because such sounds, especially when preceding vowels, are produced in explosion (Roach, 2009) and the released air is perceived as speech by the tool. What is more, though such cases may not be very frequent, researchers could be required to make their own decisions when identifying precise boundaries of pauses or utterances. Taken together, such issues highlight the importance of doing multiple listening and manual checks when using the software for analysing oral data.

Another methodological implication of the research relates to coding data for AS-units. AS-units are widely used in fluency research as a valid and reliable unit of analysis for oral data, and thus was also used in the present research. Yet, as pointed out in the previous chapter, one of my observations from the data analysis was that the majority of participants exhibited an interesting pattern in their pausing behaviour regarding the pause location. In both L1 and L2, they often tended to pause after conjunctions (e.g. 'and', 'and then' or 'because' in English or 've', 've sonra' in Turkish). It was possible that conceptualization-related pauses occurred at idea unit boundaries (to convey the content such as describing an event), rather than AS-unit boundaries (Ellis & Barkhauizen, 2005). AS-unit analysis required me to treat such pauses as part of mid-clause pausing even though they seemed to be inherently end-clause pauses (i.e. those for conceptualization reasons). Clearly, if all of such pauses had been considered as end-clause position pauses in the present study, the results would have changed. In a similar vein, Kilic and Bada (2019) showed that Turkish learners of English paused for longer following conjunctions 'because' and 'whereas' than they did preceding them. Whereas L1 English speakers were reported to have paused for significantly longer durations preceding these conjunctions. When taken together, the findings imply that the unit of analysis selected for data coding could be sensitive to cross-linguistic affects. This also suggests that fluency analysis in different languages might lead to different findings for mid-clause and end-clause pausing, and equally important a comparison of the findings across studies could be misleading.

9.4.3.Pedagogical implications

The present findings are also of interest to pedagogical practices such as language teaching and language assessment. Fluency has been one of the most important descriptors of speaking proficiency in several widely-accepted language benchmarks (e.g. ACTEFL, 2014; CEFR, 2001), and has been featured as one of the assessment criteria in rating scales of most high-stake standardized language tests (e.g. APTIS, IELTS, TOEFL). In these tests and language benchmarks, characteristics such as ‘ natural, smooth flow of language’ is highlighted (Council of Europe, 2011, p. 28-29) and frequent uses of pauses, hesitations or reformulations are referred as indicators of disfluencies. The current findings, however, demonstrate that such features in speech could be reflecting speakers’ L1 fluency behaviour, at least to some degree, rather than their disfluency. Given that such high-stake tests have important consequences for test takers, it would be worthwhile to consider refining and/or revising the assessment criteria used in these tests in a way that takes into account L1 fluency behaviour when assessing L2 fluency. One way could be to exclude those fluency measures which have been shown to reflect L1 personal speaking styles from the fluency assessment criteria.

This research has also generated findings that are of potential benefit to L2 teaching practices, e.g. in instructional settings such as schools or universities or in different learning contexts such as study-abroad or EFL contexts. In language teaching, assessment of learners’ progress is often based on fluency features in their speech. In fact, L2 teachers might place an even bigger emphasis on speech fluency in their classroom assessment because they usually equate fluency with general speaking ability or language proficiency (Tavakoli & Hunter, 2018) and typically attribute fluency indices (such as pauses, corrections, hesitations) to learners' interlanguage development. Yet, as the findings suggest, these features could be a function of learners' L1 personal speaking styles. Given the crucial role of teachers' understanding of L2 speech fluency in their classroom practices, this is an important point which could be considered in teacher training programmes as well. It would be advisable to provide L2 teachers a fluency training to raise their awareness of common fluency characteristics of L2 learners as well as of the potential impact of L1 backgrounds in learners' fluency performance. Teachers' classroom practices, be it for teaching or assessment purposes, would hugely benefit from a better understanding of the relationship between L1 and L2 fluency behaviours.

9.5. Limitations and directions for future research

While every effort has been made to ensure the reliability of the findings in Study 1 and Study 2, I should also acknowledge some of the limitations of these studies. Firstly, although both studies have yielded some important findings about the fluency relationships between L1 Turkish and L2 English, further research need to be carried out in order to validate the findings. It would be an essential next step for future studies to look into performances across more language pairs, including both structurally distant and similar languages. The current research has not focused on cross-linguistic effects; yet, in extensive cross-linguistic studies, the effects of both L1 personal speaking styles and cross-linguistic similarities and differences could be examined together and further and this would help us establish a greater degree of confidence on the issue of trait vs language specific characteristics of speech. In addition, although both studies here had a relatively decent number of participants, it is recommended to recruit more participants to increase the number of participants in each LP group; in this way, both the reliability and the generalisability of the findings could be enhanced.

This research has used only one type of task, i.e. narrative tasks, to elicit speech samples from the participants. As explained before, the choice of the tasks was based on a number of criteria and was well-informed by the relevant research. However, it is necessary to point out that although narrative tasks are predominantly used in research investigating L2 speech performance, they are one type of monologues, which involve only one speaker. In narrative tasks, there is no interaction with other speakers and we cannot observe many features of speech which are typically evident in dialogues, such as turn taking, backchannelling, interruptions, etc. (Hunter, 2017; Tavakoli, 2016; Van Os, De Jong & Bosker, 2020; Witton-Davies, 2014). In these regards, it can be said that monologic tasks do not represent the most frequently used mode of speech. In addition, research studies provide evidence suggesting that dialogic tasks encourage more fluent speech than monologic tasks (e.g. Michel, Kuiken & Vedder, 2007; Sato, 2014; Tavakoli, 2016). This means that if we are to gain a proper understanding of the relationships between L1 and L2 fluency behaviours, we need to explore fluency performances in dialogic tasks (e.g. a conversation or an interview) as well. This would be another fruitful area for further work.

Furthermore, it is important to bear in mind that any factors that influence L2 fluency development or performance could potentially exert an impact on the strength of L1-L2 fluency links. While this research was limited to exploring only a few of such factors (i.e. individual learner variables of L2 proficiency, WMC, LoR or an external variable of task structure) for reasons of time and scope, future studies could consider examining the influence of other factors (e.g. other design features such as storyline complexity) on L1-L2 associations. At this moment, it is also necessary to note that individual learner variables such as L2 proficiency are often dynamic, meaning that they change over time (Huensch & Tracy-Ventura, 2017). Longitudinal research in this regard would be of great help in understanding whether L1-L2 fluency relationships change over a period of time across the same groups of L2 learners.

Another limitation of the current research is that it was built on a quantitative approach only to investigating L1-L2 fluency associations. Fluency relationships between languages could be more complex than they at first seem; therefore, a qualitative approach to examining fluency performance in both languages would shed more light on the links between them. A qualitative approach could include 1) an in-depth discourse analysis of speech samples in both L1 and L2 for fluency indices such as such pauses preceding or following specific adverbs or conjunctions (e.g. then, and then, because, etc), or 2) retrospective interviews with participants to explore the reasons behind their pausing or repair patterns. These would lead to a more in-depth understanding of L1 and L2 speech processing and could also be a useful way of exploring different functions of pauses (filled or silent) or repair instances in speech across different LP groups.

Finally, although the overriding focus of both Study 1 and Study 2 was on the examination of L1-L2 fluency links, Study 2 was also interested in exploring the relationship between L1 and L2 lexical complexity behaviours. Two major limitations in the investigation of lexical complexity performance should be acknowledged; 1) lexical complexity was represented in a rather limited way in the study, i.e. through LD aspect only, and 2) only two measures were employed to explore the LD, i.e. TTR values and D scores. As discussed in Section 5.6, this was because 1) to the best of knowledge, there is currently no standard way of measuring lexical complexity in structurally different languages (i.e. Turkish and English in the present case), and 2) not all lexis measures are equally applicable across different languages. These are unfortunately limitations for most cross-linguistic studies examining lexical complexity, and this study has been no exception. Although the current findings suggest a link between L1 and L2 lexical complexity behaviours to a very small extent, further work is required on the investigation of other aspects, i.e. of LS and lexical density using more measures. Only then, would it be possible to draw dependable conclusions about the links between L1 and L2 lexical behaviour.

9.6. Final remarks

This thesis began with the main question of whether there is a relationship between L1 and L2 fluency behaviours, and if so, whether individual learner variables (L2 proficiency, WMC and LoR) or task structure as an external factor have a mediating role on this relationship. The suggestions of the literature as well as my own personal observations of people’s L1 and L2 speech led me to investigate the possible L1-L2 fluency links. I designed two separate but inter-related studies for this purpose; Study 1 and Study 2. Both studies have been the first to examine the fluency links between L1 Turkish and L2 English, and the effects of mediating factors, which I mentioned above, on these links. While the main focus in both studies was on examining speech fluency in

Turkish and English, a small part of Study 2 was also interested in examining L1-L2 lexical complexity relationships and the links between two aspects of the performance, i.e. fluency and lexical complexity, in each language separately.

The findings from both studies have suggested that L1 and L2 fluency behaviours are related to each other, at least to some extent. It seems that L2 breakdown and repair fluency could partly reflect L1 styles for learners in both EFL and study-abroad contexts, while speed fluency becomes associated with L1 for leaners at higher levels of LP in a study-abroad context. The L1-L2 relationships overall persist across different LP levels, and the results indicate no impact of L2 proficiency on the strength of the fluency links. Nor was there an impact of other factors, i.e. task structure, WMC and LoR. This shows that the fluency relationships are not affected by variations in any of these variables.

This research has shown that personal speaking styles of people, e.g. their tendency to pause within speech or their preference of speech rate, are carried over their L2 speech to a certain degree. This supports the suggestions of previous studies (e.g. De Jong et., 2015; Derwing, 2017; Derwing et al., 2009; Segalowitz, 2010) in that L2 fluency is partly a characteristic of individuals. That is to say, if speakers have a tendency to speak fast in their L1, they are more likely to speak fast in their L2 as well; therefore, it would be unrealistic to expect an L1 slower speaker to speak fast in their L2. This is an important point to be considered in any judgements made upon L2 fluency whether for research, testing or teaching purposes. Since fluency performance does not seem to be purely specific to a given L2, any conclusions drawn about L2 fluency could clearly be misleading unless speakers' L1 backgrounds are taken into consideration. However, as emphasized several times before, more research needs to be conducted to continue to explore the relationships between L1 and L2 fluency behaviours in order to confirm the findings, especially for different language pairs.

List of References

ACTFL (2014). American Council on the Teaching of Foreign Languages. Retrieved 1 November 2014 from: http://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012

Ahmadian, M. J. (2012a). Task repetition in ELT. *ELT Journal, 66*(3), 380-382.

Ahmadian, M. J. (2012b). The relationship between working memory capacity and L2 oral performance under task-based careful online planning condition. *TESOL Quarterly, 46*(1), 165-175.

Ahmadian, M. J. (2013). Working memory and task repetition in second language oral production. *Asian Journal of English Language Teaching, 23*(1), 37-55.

Ahmadian, M. J., & Tavakoli, M. (2011). The effects of simultaneous use of careful online planning and task repetition on accuracy, complexity, and fluency in EFL learners' oral production. *Language Teaching Research, 15*(1), 35-59.

Ahmadian, M. J., Tavakoli, M., & Vahid Dastjerdi, H. (2015). The combined effects of online planning and task structure on complexity, accuracy and fluency of L2 speech. *The Language Learning Journal, 43*(1), 41-56.

Akakura, M. (2012). Evaluating the effectiveness of explicit instruction on implicit and explicit L2 knowledge. *Language Teaching Research, 16*(1), 9-37.

Aksan, Y., Aksan, M., Koltuksuz, A., Sezer, T., Mersinli, Ü., Demirhan, U. U., . . . Yildiz, I. (2012). *Construction of the Turkish National Corpus (TNC).* Paper presented at the LREC.

Alan, D. (2004). Oxford Placement Test. In *University of Cambridge local examination syndicate*. Oxford: Oxford University Press.

Albert, A., & Kormos, J. (2011). Creativity and narrative task performance: An exploratory study. *Language Learning, 61*, 73-99.

Alptekin, C., & Erçetin, G. (2009). Assessing the relationship of working memory to L2 reading: Does the nature of comprehension process and reading span task make a difference? *System, 37*(4), 627-639.

Alptekin, C., & Erçetin, G. (2011). Effects of Working Memory Capacity and Content Familiarity on Literal and Inferential Comprehension in L2 Reading. *TESOL Quarterly,*

Awwad, A. (2017). *The effects of task complexity manipulated by intentional reasoning demands on second language learners' speech performance: interaction with language proficiency and working memory* University of Reading, Awwad, A., & Tavakoli, P. (2019). Task complexity, language proficiency and working memory: Interaction effects on second language speech performance. In *International Review of Applied Linguistics in Language Teaching* (Vol. 0).

Awwad, A., Tavakoli, P., & Wright, C. (2017). "I think that's what he's doing": Effects of intentional reasoning on second language (L2) speech performance. *System, 67*, 158-169.

Baddeley, A. (1974). *Working memory. The psychology of learning and motivation*. New York: NY Academicp.

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in cognitive sciences, 4*(11), 417-423.

Baddeley, A. (2003). Working memory: looking back and looking forward. *Nature Reviews. Neuroscience, 4*(10), 829-829.

Baddeley, A. (2015). Working memory in second language learning. In Z. Wen, M. Mota, & A. McNeill (Eds.), *Working memory in second language acquisition and processing* (pp. 17-28). Bristol: Multilingual Matters.

Baddeley, A. (2017). *Exploring working memory*. London: Routledge.

Baddeley, A., Hitch, G. J., & Allen, R. J. (2009). Working memory and binding in sentence recall. *Journal of Memory and Language, 61*(3), 438-456.

Baker-Smemoe, W., Dewey, D. P., Bown, J., & Martinsen, R. A. (2014). Does measuring L2 utterance fluency equal measuring overall L2 proficiency? Evidence from five languages. *Foreign Language Annals, 47*(4), 707-728.

Bardel, C., Gudmundson, A., & Lindqvist, C. (2012). Aspects of lexical sophistication in advanced learners' oral production: Vocabulary Acquisition and Use in L2 French and Italian. *Studies in Second Language Acquisition, 34*(2), 269-290.

Berman, R., & Verhoeven, L. (2002). Cross-linguistic perspectives on the development of text-production abilities: Speech and writing. *Written Language & Literacy, 5*(1), 1-43.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*: O'Reilly Media Inc.

Boersma, P., & Weenink, D. (2013). Praat: doing phonetics by computer (Version 5.3.04). Retrieved from http://www.praat.org/

Bolibaugh, C., & Foster, P. (2013). Memory-based aptitude for nativelike selection: The role of phonological short-term memory. In G. Granena & M. Long (Eds.), *Senstive periods, language aptitude and ultimate L2 attainment* (pp. 203-228). Amsterdam: John Benjamins.

Bosker, H. R., Pinget, A.-F., Quené, H., Sanders, T., & De Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing, 30*(2), 159-175.

Bradlow, A. R., Kim, M., & Blasingame, M. (2017). Language-indepedent talker-specificity in first-language and second-language speech production by bilingual talkers: L1 speaking rate predicts L2 speaking rate. *The Journal of the Acoustical Society of America, 141*(886).

Bulte, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: complexity, accuracy and fluency in SLA.* (Vol. 32, pp. 317). Amsterdam: John Benjamins Publishing Company.

Bulte, B., Housen, A., Pierrard, M., & Van Daele, S. (2008). Investigating lexical proficiency development over time: The case of Dutch-speaking learners of French in Brussels. *Journal of French Language Studies, 18*(3), 277-298.

Bygate, M. (2001). Effects of task repetition on the structure and control of oral language. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning , teaching and testing* (pp. 23-48). Harlow: Longman.

Carroll, J. B. (1981). Twenty-five years of research on foreign language aptitude. In K. C. Diller (Ed.), *Individual differences and universals in language learning aptitude* (pp. 83-158).

Chafe, W. L. (1980). Some reasons for hesitating. In H. W. Dechert & M. Raupach (Eds.), *Temporal variables in speech: Studies in honour of Frieda Goldman-Eisler* (pp. 271-286). The Netherlands: Mouton.

Chafe, W. L. (1994). *Discourse, consciousness and time: The flow and displacement of conscious experience in speaking and writing*. Chicago: University of Chicago Press.

Cobb, T. (2008). Web Vocabprofiler (Version 2.6) [Computer Software]. Retrieved from http://www.lextutor.ca/vp

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.). Mahwah, NJ: Erlbaum.*

Cohen, L., Manion, L., & Morrison, K. (2018). *Research methods in education (eight edition). New York: Routledge Editors.*

Conklin, K., & Schmitt, N. (2008). Formulaic Sequences: Are They Processed More Quickly than Nonformulaic Language by Native and Nonnative Speakers? *Applied Linguistics, 29*(1), 72-89.

Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. (2001). *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge University Press.

Crossley, S. A., Cobb, T., & McNamara, D. S. (2013). Comparing count-based and band-based indices of word frequency: Implications for active vocabulary research and pedagogical applications. *System, 41*(4), 965-981.

Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *Journal of the Acoustical Society of Ameirca, 111*, 2862-2873.

Cucchiarini, C., Strik, H., & Boyes, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *Journal of the Acoustical Society of America*, *107*(2), 989-999.

Daller, H., Milton, J., & Treffers-Daller, J. (2007). *Modelling and assessing vocabulary knowledge*: Cambridge University Press Cambridge.

Daller, H., Van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied linguistics, 24*(2), 197-222.

Daller, H., & Xue, H. (2007). Lexical richness and the oral proficiency of Chinese EFL students: A comparison of different measures. In M. H. Daller, J. Milton, & T.-D. Jeannine (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 150-164). Cambridge: Cambridge University Press.

Daneman, M. (1991). Working memory as a predictor of verbal fluency. *Journal of Psycholinguistic research, 20*(6), 445-464.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Memory and Language, 19*(4), 450.

Davies, A. (2003). *The native speaker: Myth and reality* (Vol. 38): Multilingual Matters.

De Bot, K. (1982). A bilingual production model: Levelt's speaking model adapted. *Applied Linguistics, 13*(1), 1-24.

De Bot, K. (1992). A bilingual production model: Levelt's speaking model adapted. *Applied Linguistics*, *13*(1), 1-24.

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.

De Jong, N. H. (2016). Predicting pauses in L1 and L2 speech: the effects of utterance boundaries and word frequency. In *International Review of Applied Linguistics in Language Teaching* (Vol. 54, pp. 113).

De Jong, N. H., & Bosker, H. R. (2013). *Choosing a threshold for silent pauses to measure second language fluency*. Paper presented at the 6th workshop on disfluency in spontaneous speech (diss).

De Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. H. (2015). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics, 36*(2), 223-243.

De Jong, N. H., & Mora, J. C. (2019). Does having good articulatory skills lead to more fluent speech in first and second languages? *Studies in Second Language Acquisition*, 1-13.

De Jong, N. H., & Perfetti, C. A. (2011). Fluency training in the ESL classroom: An experimental study of fluency development and proceduralization. *Language learning, 61*(2), 533-568.

De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. H. (2012a). Facets of speaking proficiency. *Studies in Second Language Acquisition, 34*(1), 5-34.

De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. H. (2012b). The effect of task complexity on functional adequacy, fluency and lexical diversity in speaking performances of native and non-native speakers. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 121-142).

De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. H. (2013). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics, 34*(5), 893-916.

De Jong, N. H., & Vercellotti, M. L. (2016). Similar prompts may not be similar in the performance they elicit: Examining fluency, complexity, accuracy, and lexis in narratives from five picture prompts. *Language Teaching Research, 20*(3), 387-404.

De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable buclei and measure speech rate automatically. *Behavior research methods, 41*(2), 385-390.

Declerck, M., & Kormos, J. (2012). The effect of dual task demands and proficiency on second language speech production. *Bilingualism: Language and Cognition, 15*(04), 782-796.

DeKeyser, R. (1998). Beyond focus on form: Cognitive perspectives on learning and practicing second language grammar. *Focus on form in classroom second language acquisition, 28*, 42-63.

DeKeyser, R. (2003). Implicit and explicit learning. In C. J. Doughty & M. Long (Eds.), *Handbook of Second Language Acquisition* (pp. 313-348). Oxford: Blackwell.

DeKeyser, R. (2015). Skill acquisition theory. In B. VanPatten & S. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 94-112). London: Routledge.

DeKeyser, R. (2017). Knowledge and skill in ISLA. In S. Loewen & M. Sato (Eds.), *The routledge handbook of instructed second language acquisition* (pp. 15-32).

Derwing, T. M. (2017). L2 fluency development. In S. Loewen & M. Sato (Eds.), *The routledge handbook of instructed second language acquisition* (pp. 246-259): Routhledge.

Derwing, T. M., & Munro, M. J. (2013). The development of L2 oral language skills in two L1 groups: A 7-year study. *Language Learning, 63*(2), 163-185.

Derwing, T. M., Munro, M. J., & Thomson, R. I. (2008). A longitudinal study of ESL learners' fluency and comprehensibility development. *Applied Linguistics*, *29*(3), 359-380.

Derwing, T. M., Munro, M. J., Thomson, R. I., & Rossiter, M. J. (2009). The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition, 31*(4), 533-557.

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language learning, 54*(4), 655-679.

Dewaele, J.-M., & Pavlenko, A. (2003). Productivity and lexical diversity in native and non-native speech: A study of cross-cultural effects. *Effects of the second language on the first, 3*, 120.

Dewey, D. P. (2008). Japanese vocabulary acquisition by learners in three contexts. *Frontiers: The Interdisciplinary Journal of Study Abroad, 15*, 127-148.

Di Silvio, F., Diao, W., & Donovan, A. (2016). The development of L2 fluency during study abroad: A cross-language study. *The Modern Language Journal, 100*(3), 610-624.

Díaz-Campos, M. (2004). Context of learning in the acquisition of Spanish second language phonology. *Studies in second language acquisition, 26*(2), 249-273.

Dörnyei, Z. (2007). Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies. In: Oxford: Oxford University Press.

Dörnyei, Z. (2014). *The psychology of the language learner: Individual differences in second language acquisition*: Routledge.

Dörnyei, Z., & Kormos, J. (1998). Problem-solving mechanisms in L2 communication: A psycholinguistic perspective. *Studies in second language acquisition, 20*(3), 349-385.

Du, H. (2013). The development of Chinese fluency during study abroad in China. *The Modern Language Journal, 97*(1), 131-143.

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *IRAL-International Review of Applied Linguistics in Language Teaching, 47*(2), 157-177.

Elder, C., & Ellis, R. (2009). Implicit and explicit knowledge of an L2 and language proficiency. In: Multilingual Matters.

Ellis, N. C. (1994). *Implicit and explicit learning of languages*. London: Academic Press.

Ellis, N. C. (1996). Working memory in the acquisition of vocabulary and syntax: Putting language in good order. *The Quarterly Journal of Experimental Psychology Section A, 49*(1), 234-250.

Ellis, N. C. (2005). At the interface: Dynamic interactions of explicit and implicit language knowledge. *Studies in second language acquisition, 27*(2), 305-352.

Ellis, N. C. (2006). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied linguistics, 27*(2), 164-194.

Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics, 32*, 17-44.

Ellis, N. C., & Larsen-Freeman, D. (2006). Language emergence: Implications for applied linguistics—Introduction to the special issue. *Applied linguistics, 27*(4), 558-589.

Ellis, R. (1994). A theory of instructed second language acquisition. *Implicit and explicit learning of languages*, 79-114.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in second language acquisition, 27*(2), 141-172.

Ellis, R. (2008). *The study of second language acquisition (2nd ed)*: Oxford University Press.

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied linguistics, 27*(3), 464-491.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods, 41*(4), 1149-1160.

Fiedman, N. P., & Miyake, A. (2004). The reading span test and its predictive power for reading comprehension ability. *Journal of memory and language, 51*(1), 136-158.

Fillmore, C. J. (1979). On fluency. In C. J. Fillmore, D. Kempler, & W. S. Wang (Eds.), *Individual differences in language ability and language behaviour* (pp. 85-101). New York: Academic Press.

Finardi, K., & Weissheimer, J. (2009). On the relationship between working memory capacity and L2 speech development. *Signotica, 20*(2), 367-391.

Flege, J. E., & Liu, S. (2001). The effect of experience on adults' acquisition of a second language. *Studies in second language acquisition, 23*(4), 527-552.

Foster, P. (2013). Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 85-104): Routledge.

Foster, P., Bolibaugh, C., & Kotula, A. (2014). Knowledge of nativelike selections in an L2: the influence of exposure, memory, age of onset and motivation in foreign language and immersion settings. *Studies in second language acquisition, 36*(1), 101-132.

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in second language acquisition, 18*(3), 229-323.

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

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: a unit for all reasons. *Applied Linguistics, 21*(3), 354-375.

Freed, B. F. (1995). What makes us think that students who study abroad become fluent? In B. F. Freed (Ed.), *Second language acquisition in a study abroad context*. Amsterdam: John Benjamins.

Freed, B. F. (2000). Is fluency, like beauty, in the eyes (and ears) of the beholder? In H. Riggenbach (Ed.), *Perspectives on fluency*. Michigan: The University of Michigan Press.

Freed, B. F., Segalowitz, N., & Dewey, D. P. (2004). Context of Learning and second language fluency in French: Comparing regular classroom, study abroad, and intensive domestic immersion programs. *Studies in Second Language Acquisition, 26*(2), 275-301.

French, L. M., & O'Brien, I. (2008). Phonological memory and children's second language grammar learning. *Applied Psycholinguistics, 29*(3), 463-487.

Fulcher, G. (1996). Testing tasks: Issues in task design and the group oral. *Language Testing, 13*(1), 23-51.

Gaillard, S., & Tremblay, A. (2016). Linguistic proficiency assessment in second language acquisition research: The elicited imitation task. *Language Learning, 66*(2), 419-447.

Gathercole, S. E. (2007). Working memory: a system for learning. In R. K. Wagner, A. E. Muse, & K. R. Tannenbaum (Eds.), *Vocabulary acquisition: implications for reading comprehension* (pp. 233-248): Guilford Publications.

Gathercole, S. E., Pickering, S. J., Hall, M., & Peaker, S. M. (2001). Dissociable lexical and phonological influences on serial recognition and serial recall. *The Quarterly Journal of Experimental Psychology Section A, 54*(1), 1-30.

Georgiadou, E., & Roehr-Brackin, K. (2017). Investigating executive working memory and phonological short-term memory in relation to fluency and self-repair behaviour in L2 speech. *Journal of psycholinguistic research, 46*(4), 877-895.

Gilabert, R. (2005). *Task complexity and L2 narrative oral production.* Universitat de Barcelona,

Gilabert, R. (2007). Effects of manipulating task complexity on self-repairs during L2 oral production. *International Review of Applied Linguistics in Language Teaching, 45*(3), 215-240.

Gilabert, R., & Munoz, C. (2010). Differences in attainment and performance in a foreign language: the role of working memory capacity. *International journal of English studies, 10*(1), 9-42.

Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing, 27*(3), 379-399.

Gonzalez-Fernandez, B., & Schmitt, N. (2017). Vocabulary acquisition. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition,* (pp. 280-298): Taylor Francis.

Goo, J. (2010). Working memory and reactivity. *Language Learning, 60*(4), 712-752.

Guara-Tavares, M. G. (2008). *Pre-task planning, working memory capacity and L2 speech performance.* Universidade Federal de Santa Catarina,

Guiraud, P. (1954). *Les Characteres Statistiques du Vocabularie. Essai de methodologie*. Paris: Presses Universitaires de France.

Gyllstad, H. (2013). Looking at L2 vocabulary knowledge dimensions from an assessment perspective—challenges and potential solutions. *Bardel, C., Lindqvist, C., & Laufer, B.(Eds.) L2 vocabulary acquisition, knowledge and use*: *New perspectives on assessment and corpus analysis 2*, 11-28.

Harrington, M., & Sawyer, M. (1992). L2 working memory capacity and L2 reading skill. *Studies in Second Language Acquisition, 14*, 25-38.

Hernandez, T. A. (2010). The relationship among motivation, interaction and the development of second language oral proficiency in a study-abroad context. *The Modern Language Journal, 94*(4), 600-617.

Hilton, H. (2008). The link between vocabulary knowledge and spoken L2 fluency. *Language Learning Journal, 36*(2), 153-166.

Hilton, H. (2014). Oral fluency and spoken proficiency: considerations for research and testing. In D. Leclercq, A. Edmonds, & H. Hilton (Eds.), *Measuring L2 proficiency: perspectives from SLA* (Vol. 78, pp. 27-53): Multilingual Matters.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied linguistics, 30*(4), 461-473.

Housen, A., Kuiken, F., & Vedder, I. (2012). *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (Vol. 32): John Benjamins Publishing.

Huensch, A., & Tracy-Ventura, N. (2016). Understanding second language fluency behaviour: The effects of individual differences in first language fluency, cross-linguistic differences and proficiency over time. *Applied Psycholinguistics, 38*(4), 1-31.

Huensch, A., & Tracy-Ventura, N. (2017). L2 utterance fluency development before, during and after residence abroad: A multidimensional investigation. *The Modern Language Journal, 101*(2), 275-293.

Hulstijn, J. H. (2002). Towards a unified account of the representation, processing and acquisition of second language learning. *Second Language Research, 18*(3), 193-223.

Hulstijn, J. H. (2005). Theoretical and empirical issues in the study of implicit and explicit second language learning: Introduction. *Studies in Second Language Acquisition, 27*(02), 129-140.

Hulstijn, J. H. (2011). Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second language assessment. *Language Assessment Quarterly, 8*(3), 229-249.

Hulstijn, J. H. (2012). The construct of language proficiency in the study of bilingualism from a cognitive perspective. *Bilingualism: Language and Cognition, 15*(2), 422-433.

Hulstijn, J. H. (2015). *Language proficiency in native and non-native speakers: Theory and reserch* (Vol. 41). Amsterdam: John Benjamins Publishing Company.

Hummel, K. M. (2009). Aptitude, phonological memory and second language proficiency in nonnovice adult learners. *Applied Psycholinguistics*, *30*(2), 225-249.

Hunt, K. W. (1970). Syntactic maturity in schoolchildren and adults. Monographs of the society for research in child development, 35(1), 1-67.

Hunter, A. M. (2017). *Fluency development in the ESL classroom: The impact of immediate task repetition and procedural repetition on learners' oral fluency.* St. Mary's University, Twickenham, London.

Iwashita, N. (2001). The effect of learner proficiency on interactional moves and modified output in nonnative-nonnative interaction in Japanese as a foreign language. *System, 29*(2), 267-287.

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

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Learning, 19*(1), 57-84.

Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning, 63*(1), 87-106.

Johansson, V. (2009). Lexical diversity and lexical density in speech and writing: A developmental perspective. *Lund Working Papers in Linguistics, 53*, 61-79.

Juffs, A. (2006). Working memory, second language acquisition and low-educated second language and literacy learners. *LOT Occasional Series, 6,* 89-104.

Juffs, A., & Harrington, M. (2011). Aspects of working memory in L2 learning. *Language Teaching, 44,* 137-166.

Kahng, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 English speakers: Temporal measures and stimulated recall. *Language Learning, 64*(4), 809-854.

Kahng, J. (2020). Explaining second language utterance fluency: Contribution of cognitive fluency and first language utterance fluency. *Applied Psycholinguistics*, 1-24.

Kilic, M., & Bada, E. (2019). Pausing patterns preceding and following adverbial clause conjunctions in English. *TESL-EJ, 24*(2).

Kim, Y., Dewey, D. P., Baker-Smemoe, W., Ring, S., Westover, A., & Eggett, D. L. (2015). L2 development during study abroad in China. *System, 55*, 123-133.

Kim, Y., Tracy-Ventura, N., & Jung, Y. (2016). A Measure of Proficiency or Short-Term Memory? Validation of an Elicited Imitation Test for SLA Research. *The Modern Language Journal, 100*(3), 655-673.

Kinginger, C. (2011). Enhancing language learning in study abroad. *Annual Review of applied Linguistics, 31*, 58-73.

Koponen, M., & Riggenbach, H. (2000). Overview: Varying perspectives on fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 5-24): University of Michigan.

Kormos, J. (1998). *Slef-repairs in the speech of Hungarian learners of English.* Eotvos University, Budapest.

Kormos, J. (2006). *Speech production and second language acquisition*. Mahwah, New Jersey: Lawrence Erlbaum.

Kormos, J. (2011). Speech production and the Cognition Hypothesis. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance* (Vol. 2, pp. 39-60). Amsterdam: John Benjamins.

Kormos, J., & Denes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System, 32*(2), 145-164.

Kormos, J., & Safar, A. (2008). Phonological short-term memory, working memory and foreign language performance in intensive language learning. *Bilingualism: Language and Cognition, 11*(2), 261-271.

Kormos, J., & Trebits, A. (2011). Working memory capacity and narrative task performance. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance* (Vol. 2, pp. 267-285). Amsterdam: John Benjamins.

Kormos, J., & Trebits, A. (2012). The role of task complexity, modality and aptitude in narrative task performance. *Language Learning, 62*(2), 439-472.

Krashen, S. D. (1985). *The input hypothesis*. London/NewYork: Longman.

Kuiken, F., & Vedder, I. (2007). Task complexity and measures of linguistic performance in L2 writing. *IRAL-International Review of Applied Linguistics in Language Teaching, 45*(3), 261-284.

Kuiken, F., & Vedder, I. (2008). Cognitive task complexity and written output in Italian and French as a foreign language. *Journal of Second Language Writing, 17*(1), 48-60.

Lahmann, C., Steinkrauss, R., & Schmid, M. S. (2019). Measuring linguistic complexity in long-term L2 speakers of English and L1 attriters of German. *International Journal of Applied Linguistics, 29*(2), 173-191.

Lambert, C., & Kormos, J. (2014). Complexity, accuracy and fluency in task-based L2 research: Toward more developmentally based measures of second language acquisition. *Applied Linguistics, 35*(5).

Lambert, C., Kormos, J., & Minn, D. (2017). Task repetition and second language speech processing. *Studies in Second Language Acquisition, 39*(1), 167-196.

Lancashire, I., & Hirst, G. (2009). *Vocabulary changes in Agatha Christie;s mysteries as a indication of dementia: A case study*. Paper presented at the 19th Annual Rotman Research Institute Conference, Cogitive Aging: Research and Practice, Toronto.

Laufer, B. (2003). Vocabulary acquisition in a second language: Do learners really acquire most vocabulary by reading? Some empirical evidence. *Canadian Modern Language Review, 59*(4), 567-587.

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing, 16*(1), 33-51.

Leclercq, D., & Edmonds, A. (2014). How to assess L2 proficiency? An overview of proficiency assessment research. In D. Leclercq, A. Edmonds, & H. Hilton (Eds.), *Measuring L2 proficiency: Perspectives from SLA* (Vol. 78, pp. 3-23): Multilingual Matters.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning, 40*(3), 387-417.

Lennon, P. (2000). the lexical element in spoken second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 25-42). Ann: Arbor: University of Michigan Press.

Leonard, K. R., & Shea, C. E. (2017). L2 speaking development during study abroad: Fluency, accuracy, complexity, and underlying cognitive factors. *The Modern Language Journal, 101*(1), 179-193.

Levelt, W. J. (1989). *Speaking from intention to articulation*. Cambridge, Mass: Massachusetts Institute of Technology Press.

Levelt, W. J. (1999). Producing spoken language: a blueprint of the speaker. In C. Brown & P. Hagort (Eds.), *Neurocognition of language* (pp. 83-122). Oxford: Oxford University Press.

Li, M., & Kirby, J. R. (2015). The effects of vocabulary breadth and depth on English reading. *Applied Linguistics, 36*(5), 611-634.

Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin & Review, 21*(4), 861-883.

Lindqvist, C. (2010). Inter-and intralingual lexical influences in advanced learners' French L3 oral production. *International Review of Applied Linguistics in Language Teaching, 48*(2), 131-157.

Lindqvist, C., Bardel, C., & Gudmundson, A. (2011). Lexical richness in the advanced learners' oral production of French and Italian L2. *International Review of Applied Linguistics in Language Teaching, 49*(3), 221-240.

Llanes, A. (2011). The many faces of study abroad: An update on the research on L2 gains emerged during a study abroad experience. *International Journal of Multilingualism*, *8*(3), 189-215.

Lowie, W., & Seton, B. (2012). *Essential statistics for applied linguistics*: Macmillan International Higher Education.

Mackey, A., & Gass, S. M. (2011). *Research methods in second language acquisition: A practical guide* (Vol. 7): John Wiley & Sons.

Mackey, A., & Sachs, R. (2012). Older learners in SLA research: A first look at working memory, feedback and L2 development. *Language Learning, 62*(3), 704-740.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analysing talk. Transcription format and programs* (Vol. 1): Psychology Press.

Malicka, A., & Levkina, M. (2012). Measuring task complexity: Does L2 proficiency matter? In A. Shehadeh & C. Coombe (Eds.), *Task-based language teaching in foreign language contexts: Research and implementation* (pp. 43-66). Amsterdam: John Benjamins.

Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing, 19*(1), 85-104.

Malvern, D., Richards, B., Chipere, N., & Duran, P. (2004). *Lexical diversity and language development*. New York: Palgrave Macmillan.

Marsden, E., & David, A. (2008). Vocabulary use during conversation: A cross-sectional study of development from year 9 to year 13 among learners of Spanish and French. *Language Learning Journal, 36*(2), 181-198.

Masuda, K. (2011). Acquiring interactional competence in a study abroad context: Japanese language learners' use of the interactional particle ne. *The Modern Language Journal, 95*(4), 519-540.

McCarthy, P. M., & Jarvis, S. (2007). Vocd: A theoretical and empirical evaluation. *Language Testing, 24*(4), 459-488.

McCarthy, P. M., & Jarvis, S. (2010). MLTD, vocd-D and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behaviour Research methods, 42*(2), 381-392.

Meara, P., & Miralpeix, I. (2007). D_Tools. Version 2.0. Lognostics (Centre for Applied Language Studies, University of Wales Swansea).

Meara, P., & Miralpeix, I. (2016). *Tools for researching vocabulary*: Multilingual Matters.

Michel, M. C., Kuiken, F., & Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *International Review of Applied Linguistics in Language Teaching, 45*(3), 241-259.

Miki, S. (2012). Working memory as a factor affecting L2 listening comprehension subskills. *Kumamoto University Departmental Bulletin Paper, 10,* 119-128.

Milton, J. (2009). *Measuring second language vocabulary acquisition* (Vol. 45): Multilingual Matters.

Milton, J. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *L2 vocabulary acquisition knowledge and use: new perspectives on assessment and corpus analysis* (Vol. 2, pp. 57-78).

Mitchell, A., Jarvis, S., O'Malley, M., & Konstantinova, I. (2015). Working memory measures and L2 proficiency. In Z. Wen, M. Mota, & A. McNeill (Eds.), *Working Memory in second language acquisition and processing*: Multilingual Matters.

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current directions in psychological science, 21*(1), 8-14.

Miyake, A., & Shah, P. (1999). *Models of working memory: Mechanisms of active maintenance and executive control*. New York: Cambridge University Press.

Mizera, G. J. (2006). *Working memory and L2 oral fluency.* University of Pittsburgh,

Mora, J. C., & Valls-Ferrer, M. (2012). Oral fluency, accuracy, and complexity in formal instruction and study abroad learning contexts. *TESOL Quarterly, 46*(4), 610-641.

Mota, M. (2003). Working memory capacity and fluency, accuracy, complexity and lexical density in L2 speech production. *Fragmentos, 24*, 69-104.

Munro, M. J., & Derwing, T. M. (2008). Segmental acquisition in adult ESL learners: A longitudinal study of vowel production. *Language Learning, 58*(3), 479-502.

Myles, F. (2012). Complexity accuracy and fluency. The role played by formulaic sequences in early interlanguage development. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA.* (Vol. 32).

Nation, P. (1990). *Teaching and learning vocabulary*. Boston: Heinle Cengage Learning.

Nation, P. (2001). *Learning vocabulary in another language*: Ernst Klett Sprachen.

Norris, J. M., & Ortega, L. (2006). *Synthesizing research on language learning and teaching* (Vol. 13): John Benjamins Publishing.

Norris, J. M., & Ortega, L. (2008). Defining and measuring SLA. In J. D. Catherine & H. L. Michael (Eds.), *The handbook of second language acquisition* (Vol. 27): John Wiley & Sons.

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.

Oehler, H. (1983). Grunf- und Aufbauwortschatz Deutsch (Deutsch-Turkisch). Munchen: Klett.

Official Website of Oxford English Testing. Accessed on 28 March 2018 at: *https://www.oxfordenglishtesting.com/uploadedfiles/buy_tests/oopt_meaning.pdf*.

Official Website of Syllable Count. Accessed on 28 March 2018 at: *http://www.syllablecount.com/*

Official Website of Hece Hesaplama. Accessed on 28 March 2018 at: *https://www.hesapla.online/hece-hesaplama/*

Oflazer, K. (2014). Turkish and its challenges for language processing. *Language resources and evaluation*, 48(4), 639-653.

Olynyk, M., d'Anglejan, A., & Sankoff, D. (1990). A quantitative and qualitative analysis of speech markers in the native and second language speech of bilinguals. In R. Scarcella, R. Andersen, & S. D. Krashen (Eds.), *Developing communicative competence in second languages*. Rowley, Mass: Newbury House.

Ongun, Z. (2018). *Bilingualism, vocabulary knowledge and nonverbal intelligence: Turkish-English bilingual children in the UK.* University of Reading, Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics, 24*(4), 492-518.

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*. Berlin, Germany: De Gruyter.

Ortega, L., Iwashita, N., Norris, J. M., & Rabie, S. (2002). *An investigation of elicited imitation tasks in cross linguistic SLA research*. Paper presented at the Second Language Research Forum.

O'Brien, I., Segalowitz, N., Collentine, J., & Freed, B. F. (2006). Phonological memory and lexical, narrative and grammatical skills in second language oral production by adult learners. *Applied Psycholinguistics, 27,* 377-402.

O'Brien, I., Segalowitz, N., Collentine, J., & Freed, B. F. (2007). Phonological memory predicts second language oral fluency gains in adults. *Studies in Second Language Acquisition, 29,* 557-582.

Pallant, J. (2013). *SPSS survival manual*. London: McGraw-Hill Education.

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics, 30*(4), 590-601.

Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research, 31*(1), 117-134.

Paradis, M. (2009). *Declarative and procedural determinants of second languages*: John Benjamins Publishing.

Pavlenko, A. (1999). New approaches to concepts in bilingual memory. *Bilingualism: Language and Cognition, 2*(3), 209-230.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. *Language and Communication*.

Peltonen, P. (2018). Exploring connections between first and second language fluency: A mixed methods approach. *The Modern Language Journal, 102*(4), 676-692.

Pica, T., Holliday, L., Lewis, N., & Morgenthaler, L. (1989). Comprehensible output as an outcome of linguistic demands on the learner. Studies in second language acquisition, 11(01), 63-90.

Plonsky, L., & Ghanbar, H. (2018). Multiple regression in L2 research: A methodological synthesis and guide to interpreting R2 values. *The Modern Language Journal, 102*(4), 713-731.

Plonsky, L., & Oswald, F. L. (2014). How Big Is “Big”? Interpreting Effect Sizes in L2 Research. *Language Learning, 64*(4), 878-912.

Prefontaine, Y., Kormos, J., & Johnson, D. E. (2016). How do utterance measures predict raters’ perceptions of fluency in French as a second language? *Language Testing, 33*(1), 53-73.

Préfontaine, Y., & Kormos, J. (2015). The relationship between task difficulty and second language fluency in French: A mixed methods approach. *The Modern Language Journal, 99*(1), 96-112.

Préfontaine, Y., & Kormos, J. (2016). A qualitative analysis of perceptions of fluency in second language French. *International Review of Applied Linguistics in Language Teaching, 54*(2), 151-169.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Riazantseva, A. (2001). Second language proficiency and pausing: A study of Russian speakers of English. *Studies in Second Language Acquisition, 23*(04), 497-526.

Roach, P. (2009). *English phonetics and phonology paperback with Audio CDs (2): A practical course*: Cambridge University Press.

Robinson, P. (2001). Task complexity, task difficulty and task production: Exploring interactions in a componential framework. *Applied Linguistics, 22*(1), 27-57.

Robinson, P. (2011). *Second language task complexity: Researching the cognition hypothesis of language learning and performance* (Vol. 2): John Benjamins Publishing.

Robson, C. (2002). *Real world research: A resource for social scientists and practitioner-researchers* (Vol. 2). Oxford: Blackwell.

Rossiter, M. J. (2009). Perceptions of L2 fluency by native and non-native speakers of English. *Canadian Modern Language Review, 65*(3), 395-412.

Révész, A. (2012). Working memory and the observed effectiveness of recasts on different L2 outcome measures. *Language Learning, 62*(1), 93-132.

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., 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.

Sato, M. (2014). Exploring the construct of interactional oral fluency: Second Language Acquisition and Language Testing approaches. *System*, *45*, 79-91.

Saito, K. (2013). Age effects on late bilingualism: The production development of of/ɹ/by high-proficiency Japanese learners of English. *Journal of Memory and Language, 69*(4), 546-562.

Saito, K. (2015). Experience effects on the development of late second language learners' oral proficiency. *Language Learning, 65*(3), 563-595.

Saito, K., & Brajot, F. X. (2013). Scrutinizing the role of length of residence and age od acquisition in the interlanguage pronunciation development of English of/ɹ/by late Japanese bilinguals. *Bilingualism: Language and Cognition, 16*(4), 847-863.

Saito, K., Ilkan, M., Magne, V., Tran, M. N., & Suzuki, S. (2018). Acoustic characteristics and learner profiles of low-, mid- and high-level second language fluency. *Applied Psycholinguistics, 39*(3), 593-617.

Schmid, M. S., & Fagersten, K. (2010). Disfluency markers in L1 attrition. *Language Learning, 60*(4), 753-791.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*: Springer.

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning, 64*(4), 913-951.

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. New York: Routledge.

Segalowitz, N. (2016). Second language fluency and its underlying cognitive and social determinants. *International Review of Applied Linguistics in Language Teaching, 54*(2), 79-95.

Segalowitz, N., French, L., & Guay, J. D. (2017). What features best characterize adult second language utterance fluency and what do they reveal about fluency gains in short-term immersion? *Canadian Journal of Applied Linguistics, 20*(2), 90-116.

Serafini, E. J. (2013). *Cognitive and psychosocial factors in the long-term development of implicit and explicit second language knowledge in adult learners of Spanish at increasing proficiency.* Georgetown University.

Serrano, R., Llanes, A., & Tragant, E. (2011). Analysing the effect of context of second language learning: Domestic intensive and semi-intensive courses vs study abroad in Europe. *system, 39*(2), 133-143.

Shao, Z., Roelofs, A., & Meyer, A. S. (2012). Sources of individual differences in the speed of naming objects and actions: The contribution of executive control. *The Quarterly Journal of Experimental Psychology, 65*(10), 1927-1944.

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics, 17*(1), 38-62.

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

Skehan, P. (2003). Task-based instruction. *Language Teaching, 36*(1), 1-14.

Skehan, P. (2008). Interlanguage and language transfer. In B. Spolsky & F. M. Hult (Eds.), *The handbook of educational linguistics*: John Wiley & Sons.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied linguistics, 30*(4), 510-532.

Skehan, P. (2014). Limited attentional capacity, second language performance, and task-based pedagogy. In P. Skehan (Ed.), *Processing perspectives on task performance* (pp. 211-260). Amsterdam, the Netherlands: John Benjamins.

Skehan, P. (2015). Limited attention capacity and cognition. In M. Bygate (Ed.), *Domains and directions in the development of TBLT: A decade of plenaries from the international conference* (pp. 123-155): John Benjamins Publishing Company.

Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language teaching research, 1*(3), 185-211.

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning, 49*(1), 93-120.

Skehan, P., & Foster, P. (2007). Complexity, accuracy, fluency and lexis in tas-based performance: A meta-analysis of the Ealing research. In S. Van Daele, A. Housen, F. Kuiken, F. Pierrard, & I. Vedder (Eds.), *Complexity, accuracy and fluency in second language use and teaching* (pp. 207-226).

Skehan, P., Foster, P., & Shum, S. (2016). Ladders and Snakes in Second Language Fluency. In *International Review of Applied Linguistics in Language Teaching* (Vol. 54, pp. 97).

Speciale, G., Ellis, N. C., & Bywater, T. (2004). Phonological sequence learning and short-term store capacity determine second language vocabulary acquisition. *Applied Psycholinguistics, 25*, 293-321.

Strömqvist, S., Johansson, V., Kriz, S., Ragnarsdottir, H., Aisenman, R., & Ravid, D. (2002). Toward a cross-linguistic comparison of lexical quanta in speech and writing. *Written Language & Literacy, 5*(1), 45-67.

Suzuki, Y., & DeKeyser, R. (2019). The interface of explicit and implicit knowledge in a second language: Insights from individual differences in cognitive aptitudes. *Language Learning, 64*(4), 747-790.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics (6th ed.)*. Boston: Pearson Education.

Tavakoli, P. (2009). Assessing L2 task performance: Understanding effects of task design. *System, 37*(3), 482-495.

Tavakoli, P. (2011). Pausing patterns: Differences between L2 learners and native speakers. *ELT Journal, 65*(1), 71-79.

Tavakoli, P. (2016). Speech fluency in monologic and dialogic task performance. *International Review of Applied Linguistics in Language Teaching, 54*(2), 131-150.

Tavakoli, P., Campbell, C., & McCormack, J. (2016). Development of speech fluency over a short period of time: Effects of pedagogic intervention. *TESOL Quarterly, 50*(2), 447-471.

Tavakoli, P., & Foster, P. (2011). Task design and second language performance: The effect of narrative type on learner output. *Language Learning, 61*, 37-72.

Tavakoli, P., & Hunter, A.-M. (2018). Is fluency being 'neglected' in the classroom? Teacher understanding of fluency and related classroom practices. *Language Teaching Research, 22*(3), 330-349.

Tavakoli, P., Nakatsuhara, F., & Hunter, A. M. (2019). Aspects of fluency across assessed levels of speaking proficiency. *The Modern Language Journal*.

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239-277). Amsterdam: John Benjamins.

Tavakoli, P., & Uchihara, T. (2019). To what extent are multiword sequences associated with oral fluency? *Language Learning*.

Thai, C., & Boers, F. (2016). Repeating a monologue under increasing time pressure: Effects on fluency, complexity and fluency. *TESOL Quarterly, 50*(2), 369-393.

Thomson, C. (2014). *Guided planning, task complexity and second language oral development.* University of Central Lancashire,

Tidball, F., & Treffers-Daller, J. (2008). Analysing lexical richness in French learner language: What frequency lists and teacher judgements can tell us about basic and advanced words. *Journal of French Language Studies*, *18*(3), 299-313.

Towell, R. (2012). Complexity, accuracy and fluency from the perspective of psycholinguistic second language acquisition research. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (Vol. 32): John Benjamins Publishing.

Towell, R., & Dewaele, J.-M. (2005). The role of psycholinguistic factors in the development of fluency amongst advanced learners of French. In J.-M. Dewaele (Ed.), *Focus on French as a foreign language* (pp. 210-239). Clevedon: Multilingual Matters.

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics, 17*(1), 84-119.

Tracy-Ventura, N., McManus, K., Norris, J. M., & Ortega, L. (2014). Repeat as much as you can: Elicited imitation as a measure of oral proficiency in L2 French. In D. Leclercq, A. Edmonds, & H. Hilton (Eds.), *Measuring L2 proficiency: Perspectives from SLA* (Vol. 78, pp. 143-166). Bristol: Multilingual Matters.

Treffers-Daller, J. (2013). Measuring lexical diversity among L2 learners of French. In S. Jarvis & M. H. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures*: John Benjamins Publishing.

Treffers-Daller, J., & Korybski, T. (2015). Using lexical diversity measures to operationalise language dominance in bilinguals. In J. Treffers-Daller & C. Silva-Corvalan (Eds.), *Language dominance in bilinguals: Issues of measurement and operationalization* (pp. 106-123). Cambridge: Cambridge University Press.

Treffers-Daller, J., Parslow, P., & Williams, S. (2018). Back to basics: How measures of lexical diversity can discriminate between CEFR levels. *Applied Linguistics, 39*(3), 302-327.

Tremblay, A. (2011). Proficiency assessment standards in second language acquisition research. *Studies in Second Language Acquisition, 33*(3), 339-372.

Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmental: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition, 28*(1), 1-30.

Ullman, M. T. (2015). The declarative/procedural model. In B. VanPatten, G. D. Keating, & S. Wulff (Eds.), *Theories in second language acquisition: An introduction* (pp. 135-158): Routledge.

Van den Noort, M. W., Bosch, P., & Hugdahl, K. (2006). Foreign language proficiency and working memory capacity. *European Psychologist, 11*(4), 289-296.

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing, 29*(3), 325-344.

Van Os, M., De Jong, N. H., & Bosker, H. R. (2020). The effect of turn-taking behaviour on perceived fluency in native and non-native speech. *Language Learning*.

Vasylets, O., Gilabert, R., & Manchon, R. M. (2017). The effects of mode and task complexity on second language production. *Language Learning, 67*(2), 394-430.

Verhagen, J. (2011). Verb placement in second language acquisition: Experimental evidence for the different behaviour of auxiliary and lexical verbs. *Applied Psycholinguistics, 32*(4), 821-858.

Vinther, T. (2002). Elicited imitation: A brief overview. *International journal of applied linguistics, 12*(1), 54-73.

Vyatkina, N. (2012). The development of second language writing complexity in groups and individuals: A longitudinal learner corpus study. *The Modern Language Journal, 96*(4), 576-598.

Wang, Z. (2014). Developing accuracy and fluency in spoken English of Chinese learners. *English Language Teaching, 7*(2), 110-118.

Weitz, D., O'Shea, G., Zook, N., & Needham, W. (2011). Working memory and sequence learning in the Hebb digits task: Awareness is predicted by individual differences in operation span. *The American Journal of Psychology, 124*(1), 49-62.

Wen, Z. (2015). Working memory in second language acquisition and processing: The phonological/executive model. In Z. Wen, M. Mota, & A. McNeill (Eds.), *Working memory in second language acquisition and processing* (pp. 41-62). Bristol: Multilingual Matters.

Wen, Z., & Li, S. (2019). Working memory in L2 learning and processing. In J. Shcwieter & A. Benati (Eds.), *The Cambridge handbook of language learning* (pp. 365-389). Cambridge: Cambridge University Press.

Wen, Z., Mota, M., & McNeill, A. (2015). *Working memory in second language acquisition and processing* (Vol. 87). Bristol: Multilingual Matters.

West, D. E. (2012). Elicited imitation as a measure of morphemic accuracy: Evidence from L2 Spanish. *Language and Cognition, 4*(3), 203-222.

Witton-Davies, G. (2014). *The study of fluency and its development in monologue and dialogue.* Lancaster University, Lancaster, UK.

Wolfe-Quintero, K., Ingaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy & complexity* (Vol. 17): University of Hawaii Press.

Wood, D. (2010). *Formulaic language and second language speech fluency: Background, evidence and classroom applications*: Bloomsbury Publishing.

Wray, A. (2002). Formulaic language in computer-supported communication: Theory meets reality. *Language Awareness, 11*(2), 114-131.

Wray, A. (2012). What do we (think we) know about formulaic language? An evaluation of the current state of play. *Annual Review of Applied Linguistics, 32*, 231-254.

Wright, C. (2010). *Role of working memory in SLA*. Saarbrucken: VDM Publishing House.

Wright, C. (2013). An investigation of working memory effects on oral grammatical accuracy and fluency in producing questions in English. *TESOL Quarterly, 47*(2), 352-374.

Wright, C. (2015). Working memory and L2 development across the lifespan: A Commentary. In Z. Wen, M. Mota, & A. McNeill (Eds.), *Working memory in second language acquisition and processing* (Vol. 87): Multilingual Matters.

Wright, C. (2018). Effects of study abroad on L2 Mandarin Chinese: Effects of time and task. In C. Sanz & A. Morales-Front (Eds.), *The Routledge handbook of study abroad research and practice* (pp. 1-17). New York: Routledge.

Wu, S. L., & Ortega, L. (2013). Measuring global oral proficiency in SLA research: A new elicited imitation test of L2 Chinese. *Foreign Language Annals, 46*(4), 680-704.

Yan, X., Maeda, Y., Lv, J., & Ginther, A. (2016). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing, 33*(4), 497-528.

Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied Linguistics, 31*(2), 236-259.

Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and online planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics, 24*(1), 1-27.

Zareva, A., Schwanenflugel, P., & Nikolova, Y. (2005). Relationship between lexical competence and language proficiency: Variable sensitivity. *Studies in Second Language Acquisition, 27*(4), 567-595.

Appendices

Appendix A. Background Questionnaire used in Study 1.

School of Literature and Languages
Department of English Language and Applied Linguistics

University of Reading

BACKGROUND QUESTIONNAIRE

Dear Participants,

The following questionnaire is part of my PhD research Project in Applied Linguistics which investigates relationship between first and second language speaking ability.

Please read and tick (X) the box.

☐ *"By completing and returning this questionnaire I understand that I am giving consent for my responses to be used for the purposes of this research project"*

Part. A. General Information

1. Age: ....
2. Gender: (a) male (b) female (c) other
3. Level of Education (your most recent educational level/ the latest degree you have finished)
4. Major: ………………………………………………………………………………....
5. Country of residence: ………………………………………………………………….
6. Country of origin: …………………………………………………………………….
7. Known languages other than English: ……………………………………………...
8. Do you have vision or hearing problems? If yes, please specify.
   ………………………………………………………………………………………….
9. Have you had ever lived or studied in another country? If yes, please state the period of time you stayed. (years, months, weeks, etc.) and the language you used
   …………………………………………………………………………………………...

Part B. Use of English

1. Age of first exposure (indicate the date you started learning or using English):

…………………………………………………………………………………………………

2. Context of first exposure (e.g. at home, with friends, at school, at work, online, etc.)

   …………………………………………………………………………………………………

3. Number of years of English instruction that you have received:

   …………………………………………………………………………………………………

4. How much do you hear or use English outside the classroom? Please consider the following

- People speaking English to you at home (parents, grandparents) ____________ (hrs)
- Relatives/ friends speaking English to you ____________ (hrs)
- Self-study English ____________ (hrs)
- Watching television in English ____________ (hrs)
- Listening to radio ____________ (hrs)
- Writing to your friends ____________(hrs)
- Others (specify) ……………………………………………………………………………

5. Have you taken any standardized language proficiency test (e.g. TOEFL, IELTS)? Please indicate the name of the test and the score you received.

   …………………………………………………………………………………………………

THANK YOU FOR YOUR TIME AND CONTRIBUTION

Appendix B. . Oxford Placement Test (Grammar Section Only) (Alan, 2004).

Oxford Placement Test 2

Grammar Test PART 1

Name ....

Total Listening .... / 100

Total Grammar .... / 100

Grand Total .... / 200

Look at these examples. The correct answer is ticked.

a In warm climates people | like | likes | are liking | sitting outside in the sun.

b If it is very hot, they sit | at | in | under | the shade.

Now the test will begin. Tick the correct answers.

1 Water | be freezing | is freezing | freezes | at a temperature of 0°C. 1 ____

2 In some countries | there is | is | it is | dark all the time in winter. 2 ____

3 In hot countries people wear light clothes | for keeping | to keep | for to keep | cool. 3 ____

4 In Madeira they have | the good | good | a good | weather almost all year. 4 ____

5 Most Mediterranean countries are | more warm | the more warm | warmer | in October than in April. 5 ____

6 Parts of Australia don't have | the | some | any | rain for long periods. 6 ____

7 In the Arctic and Antarctic | it is | there is | it has | a lot of snow. 7 ____

8 Climate is very important in | most of | most | the most | people's lives. 8 ____

9 Even now there is | little | few | less | we can do to control the weather. 9 ____

10 In the future | we'll need | we are needing | we can need | to get a lot of power from the sun and the wind. 10 ____

11 For many people the name Pelé still means | the more | the most | most | famous footballer in the world. 11 ____

12 Pelé | had been | is | was | born in 1940. 12 ____

13 His mother | not want | wasn't wanting | didn't want | him to become a footballer. 13 ____

14 But his father | made him to | made him | would make him to | practise every day. 14 ____

15 By 1956 he | has joined | joined | had joined | the Brazilian club, *Santos*, and had scored in his first game. 15 ____

subtotal /15

16 In 1957 he [has been picked | was picked | was picking] for the Brazilian national team. 16 ____

17 The next World Cup Finals were in 1958 and Pelé was looking forward to [play | playing | the play]. 17 ____

18 And [even though | even so | in spite of] he was injured he helped Brazil to win the final. 18 ____

19 Pelé was [a such | such a | a so] brilliant player that he helped Brazil win 3 World Cups. 19 ____

20 He didn't stop [playing | to play | play] for *Santos* till he was 34. 20 ____

21 After calling it a day in 1974, he came [from | off | out of] retirement and played for *New York Cosmos*. 21 ____

22 [Till | By | In] the end of his career he had scored over a thousand goals. 22 ____

23 He then settled for a role [as | like | in] a sporting ambassador for Brazil. 23 ____

24 By the end of the 20th Century he had received a great [many | number | deal] of awards. 24 ____

25 Though honoured with the title *Athlete of the Century*, he will always be remembered [as footballer | as a footballer | as the footballer]. 25 ____

Football, or soccer as it is sometimes known, [has been | is being | was] played 26 ____

for [above | over | more that] 150 years, but the first World Cup 27 ____

competition [has not been | was not | was not being] held until 1930, 28 ____

when Uruguay [could win | were winning | won] the first professional final. 29 ____

Four teams had entered from Europe, but with [a little | little | few] success. 30 ____

The 1934 World Cup was again won by [a | the | their] home team, Italy, 31 ____

[who | which | that] went on to win the 1938 final as well. Winning successive 32 ____

finals is something that [is not | was not | has not been] achieved again 33 ____

until Brazil managed [them | these | it] in 1958 and 1962. 34 ____

If Brazil [would have won | would win | had won] again in 1966 then the FIFA 35 ____

authorities would have needed to [have | let | make] the original World Cup replaced. 36 ____

However, England stopped the Brazilians [to get | getting | get] a third successive win. 37 ____

In the 1970s the honours were shared [among | between | inside] Europe and South America. 38 ____

Argentina succeeded [to win | at winning | in winning] in 1978, but in 1982, in Spain, 39 ____

they had [difficulty in | difficulties to | difficulty to] getting beyond the early stages. 40 ____

They won again in Mexico in 1986, [where | which | while] Maradonna 41 ____

managed to win [much | some | any] of the games, especially the one 42 ____

against England, almost [by his own | by himself | on himself.] The 1990s finals were 43 ____

dominated by European teams [except | apart | save] from Brazil's win in the USA in 1994, 44 ____

with the 1998 finals in France again [to be | being | having] won by the hosts. 45 ____

Throughout the 1990s police in the host countries [was | were | have been] kept busy keeping 46 ____

rival fans apart, but [there was | there were | it was] to be no such problems when the first 47 ____

World Cup Finals of the 21st century took [part | place | hold] in Japan and South Korea in 2002. 48 ____

Football's third century [has seen | saw | seeing] success for a number of footballing nations in 49 ____

Africa and Asia, who [may well | may as well | might as well] prove to be the teams of the future. 50 ____

subtotal /35

Grammar Test PART 2

51 Millions of | persons | people | peoples | around the world now use the Internet almost every day. 51 ____

52 The majority of children in the UK | have | has | are having | access to a PC. 52 ____

53 Learning to use the Internet is not the same | as | like | than | learning traditional skills. 53 ____

54 Most of us start off with email, | who | which | what | is fairly easy to use. 54 ____

55 Children generally find using computers easy, but some adults can't get used | to work | to working | work | with them. 55 ____

56 There aren't | no | any | some | shortcuts to becoming proficient – everyone needs training and practice. 56 ____

57 Those who do best are those who also use computers a lot | on their own | by their own | on themselves |. 57 ____

58 It's no use | in trying | to try | trying | to become an expert just by reading books. 58 ____

59 There are many who wish they | started | would have started | had started | learning earlier. 59 ____

60 A few unsuccessful learners have resigned themselves to never | know | knowing | known | how to use the Internet. 60 ____

61 Some new users quickly become almost addicted | to be | to being | be | on line. 61 ____

62 Others decide they would just | rather | prefer | better | not have anything to do with computers. 62 ____

63 The trend continues | to be | be | by being | for computers to get smaller and smaller. 63 ____

64 Some companies already have more palmtops | that | than | as | desktops. 64 ____

65 It is thought that we'll have mobile phones as powerful as PCs | till | by | in | the end of the decade. 65 ____

Below is a letter written to the 'advice' column of a daily newspaper. Tick the correct answers.

Dear Marge,

| I'm writing | I will write | I should write | to you because I 66 ____

| am not knowing | don't know | know not | what to do. I'm twenty-six and a teacher at 67 ____

a primary school in Norwich where | I'm working | I've worked | I work | for the last five years. 68 ____

When I | was | have been | had been | there for a couple of years, one of the older members of staff 69 ____

| would leave | left | had been leaving |, and a new teacher 70 ____

| would be | became | was | appointed to work in the same department as me. 71 ____

We | worked | have worked | should work | together with the same classes during her first year 72 ____

and had the | opportunity for building | possibilities to build | chance to build | up a good professional 73 ____

relationship. Then, about eighteen months after | she has arrived | to have arrived | arriving | 74 ____

in Norwich, she decided to buy | her own | herself | her a | house. 75 ____

subtotal /25

She was tired of | to live | live | living | in rented accommodation and wanted a place 76 ____
| by her own | of her own | of herself | . At about the same time, I 77 ____
| was given | have been given | gave | notice by the landlord of the flat 78 ____
| what I was living | that I had lived | I was living | in 79 ____
and she asked me if I | liked | had liked | would like | to live 80 ____
with her. She | said | told | explained | me that by the time she 81 ____
| would pay | would have paid | had paid | the mortgage 82 ____
and the bills | it | there | they | wouldn't be 83 ____
| a lot | many | few | left to live on. She suggested 84 ____
| us to | we should | we may | share the house and share the costs. 85 ____
It seemed like a good idea, so after | we'd agreed | we could agree | we agreed with | all the details 86 ____
| what | that | who | needed to be sorted out, we moved into the new house together. 87 ____
At the end of this month | we have lived | we have been living | we'll have been living | 88 ____
together for a year and a half. It's the first time | I live | I'm living | I've lived | with anybody before, but 89 ____
| I should guess | I might have guessed | I'd have guessed | what would happen. I've fallen in love with 90 ____
her and now she's been offered another job 200 miles away and is going to move. I don't know what to do. Please give me some advice.

Yours in shy desperation,

Steve

Look at the following examples of question tags in English. The correct form of the tag is ticked.

a He's getting the 9.15 train, | isn't he ✓ | hasn't he | wasn't he | ?

b She works in a library, | isn't she | doesn't she ✓ | doesn't he | ?

c Tom didn't tell you, | hasn't he | didn't he | did he ✓ | ?

d Someone's forgotten to switch off the gas, | didn't one | didn't they | haven't they ✓ | ?

Now tick the correct question tag in the following 10 items:

91 Steve's off to China, | has he | hasn't he | isn't he | ? 91 ____
92 It'll be a year before we see him again, | won't it | won't we | shan't it | ? 92 ____
93 I believe he's given up smoking, | isn't he | don't I | hasn't he | ? 93 ____
94 I'm next on the list to go out there, | am not I | are I | aren't I | ? 94 ____
95 No doubt you'd rather he didn't stay abroad too long, | shouldn't you | wouldn't you | hadn't you | ? 95 ____
96 He's rarely been away for this long before, | is he | hasn't he | has he | ? 96 ____
97 So you think he'll be back before November, | shall he | will he | do you | ? 97 ____
98 Nobody's disagreed with the latest proposals, | did he | has he | have they | ? 98 ____
99 We'd better not delay reading this any longer, | should we | did we | had we | ? 99 ____
100 Now's hardly the time to tell me you didn't need a test at all, | did you | is it | isn't it | ? 100 ____

subtotal /25

Appendix C. Oral narrative tasks used in Study 1 (Henry comics by Don Trachte).

Task A. The ice-skating boy

Task B. The boy looking after his car

Appendix D. EIT sentences and task instructions.

You are going to listen to a set of sentences one by one. After each sentences, you will hear a beed sound. Your task is to repeat the sentences after hearing the beed sound as exaclly as possible. You will not take any notes during the task, and will have only one attempt to repeat the sentence. If you cannot repeat a sentence, please wait until you hear the next once. Your performance will be audio recorded during the test.

EIT sentences (adopted from Ortega et al., 2002), *listed from the least number of syllables to the most as incidated in parenthesis.*

1. I have to get a haircut (7)
2. The red book is on the table (8)
3. The streets in this city are wide (8)
4. He takes a shower every morning (9)
5. What did you say you were doing today? (10)
6. I doubt that he knows how to drive that well (10)
7. After dinner I had a ling, peaceful nap (11)
8. It is possible that it will rain tomorrow (12)
9. I enjoy movies which have a happy ending (12)
10. The houses are very nice but too expensive (12)
11. The little boy whose kitten died yesterday is sad (13)
12. That restaurant is supposed to have very good food (13)
13. I want a nice, big house in which my animals can live (14)
14. You really enjoy listening to country music, don't you (14)
15. She just finished painting the inside of her apartment (14)
16. Cross the street at the light and then just continue straight ahead (15)
17. The person I'm dating has a wonderful sense of humour (15)
18. She only orders meat dishes and never eats vegetables (15/16)
19. I wish the price of town houses would become affordable (15)

20. I hope it will get warmer sooner this year than it did last year (16)

21. A good friend of mine always takes care of my neighbour’s three children (16)

22. The black cat that you fed yesterday was the one chased by the dog(16)

23. Before he can go outside, he has to finish cleaning his room (16)

24. The most fun I've ever had was when we went to the opera (16)

25. The terrible thief whom the police caught was very tall and thin (17)

26. Would you be so kind as to hand me the book which is on the table? (17)

27. The number of people who smoke cigars is increasing every year (17/18)

28. I don't know if the 11:30 train has left the station yet (18)

29. The exam wasn't nearly as difficult as you told me it would be (18)

30. There are a lot of people who don’t eat anything at all in the morning (19)

Appendix E. Ethical Approval granted for the study.

School of Literature and Languages
Department of English Language and Applied Linguistics

University of Reading

ETHICS COMMITTEE

Project Submission

Note All sections of this form must be completed.

Principal Investigator (Supervisor): Parvaneh Tavakoli

Student name: Zeynep Duran Karaoz

Department: English Language and Applied Linguistics

Title of Project: An Exploratory Study of L2 Oral Performance: Fluency and Lexical Complexity in L1 Turkish and L2 English

Proposed starting date: 31 August 2016

Number of participants that you require consent from (approximate): 100

I confirm that to the best of my knowledge the Ethics and Research Committee have been made aware of all relevant information. I undertake to inform the Committee of any such information which subsequently becomes available whether before or after the research has begun.

I confirm that a list of the names and addresses of the participants in this project will be compiled and that this, together with a copy of the Consent Form, will be retained. All copies of the Consent Forms will be submitted with a copy of the dissertation.

Signed:

Dr. Parvaneh Tavakoli (Supervisor) Date...25 August 20 [redacted]

Zeynep Duran Karaoz (Student) Date..25 August 2016

31/8/2016

[redacted]

(ethics committee)

1

Appendix F. Information Sheet given to the participants in Study 1.

School of Literature and Languages
Department of English Language and Applied Linguistics

University of Reading

Researcher:
Zeynep Duran Karaoz
Phone: [redacted]
Email: Zeynep.DuranKaraoz@pgr.reading.ac.uk

Supervisor:
Parvaneh Tavakoli
Phone: [redacted]
Email: p.tavakoli@reading.ac.uk

Department of English Language and Applied Linguistics

URS Building
The University of Reading
Whiteknights, PO Box 219
Reading RG6 6AW

Phone [redacted]

[redacted] [redacted]
Email appling@reading.ac.uk
p.a.thompson@reading.ac.uk

INFORMATION SHEET

This project aims to understand if there is a relationship between first and second language speaking ability.

You have been selected to take part in this study because you speak Turkish as your native language and English as your second language. 60 (approx.) learners, aged between 19-35, will participate in the study. Data collection procedure will be run on two separate days. On Day 1, you will be given a short questionnaire and two English language proficiency tests; one after another with short breaks in between. In two of these tests, you will be asked to answer a set of questions on paper. In the final one, you will hear a set of sentences and will repeat them after hearing each sentence. Your speech will be audio-recorded to be scored. You will be contacted to meet on a different day which is convenient for you. On Day 2, you will be met individually on a one-to-one setting, and you will be asked to look at a set of pictures and tell a short story both in Turkish and English. Your speech will be audio-recorded, transcribed and analysed.

All the data collected will be securely kept in a password-protected computer, and will be accessed only by the investigator, Zeynep Duran Karaoz, and her supervisors. The data will be used for research purposes only and will be destroyed immediately after the completion of the thesis (in five years' time). Your names will not be mentioned in the project and you will be able to withdraw from the study at any time you wish to. Your privacy and confidentiality will be carefully observed.

This project has been subject to ethical review by the School Ethics and Research Committee, and has been allowed to proceed under the exceptions procedure as outlined in paragraph 6 of the University's *Notes for Guidance* on research ethics.

If you have any queries or wish to clarify anything about the study, please feel free to contact my supervisor at the address above or by email at [include supervisor's email address here]

Signed

Appendix G. Consent form

School of Literature and Languages
Department of English Language and Applied Linguistics

University of Reading

ETHICS COMMITTEE

Consent Form

Project title: An Exploratory Study of L2 Oral Performance: Fluency and Lexical Complexity in L1 Turkish and L2 English

I understand the purpose of this research and understand what is required of me; I have read and understood the Information Sheet relating to this project, which has been explained to me by Zeynep Duran Karaoz. I agree to the arrangements described in the Information Sheet in so far as they relate to my participation.

I have received a copy of this Consent Form and of the accompanying Information Sheet.

Name:

Signed:

Date:

Appendix H. A sample of coded transcription with coding conventions used.

Turkish

| **erm 0.25** çocuğun biri **(0.52)** evinin yakınlarında bir **(0.57)** buz pateni yapmaya karar vermiş **(1.00)** | buz pateni yaparken **(0.60)** **::** buzun altında bir **(0.52)** balık görmüş **(1.00)** | iste baliği izlerken **(0.62)** **::** buz pateni yapmaya devam etmiş :: ve baliğin etrafında daireler çizmiş **(1.00)** | bir sure sonra <daireler> [/-](**0.95**) adam suya düşmüş **(0.30)** :: <daireler> [/-] **a** şey <buzda> [/-] çünkü buz **(0.90)** <za>[/-] zarar görmüş :: ve **(0.42)** buzun altında balıklarla **(0.52)** görünüyor en son |

English

| **erm 0.25** one time a guy **(0.41)** decides to ice skate **(0.58)** near **(0.45)** his home **(0.80)** | nd then while he put the shoes **(0.42)** **::** he started to skate **(0.78)** | and then under the ice he saw a fish **(1.15)** | and <the> **[///]** he skated **::** while watching the fish **(0.84)** | and drives circles around the fish **(0.55)** | <and> [/-] **erm 0.73** but **(2.19)** **erm 0.85** he keeps circling **::** so the ice gets **(1.00)** damaged **(0.50)** | **erm 0.40** after some time <the ice> [/-] **(0.88)** he **(0.32)** fall into the **(0.28)** cold water | and ice got broke **(1.01)** | and **erm 0.34** he now under the ice with some fishes |

Coding conventions used

|	AS-unit boundary	**[/]**	< repeated words>
::	clause boundary	**[//]**	<replaced words>
( )	silent pause	**[///]**	<reformulated words>
erm	filled pause	**[/-]**	<false starts>

Appendix I. Q-Q plots obtained for dependent variables (i.e fluency measures in Turkish and English) and independent variables (i.e. language proficiency scores) of Study 1.

Figure 1. Q-Q plot for Turkish number of MCFP

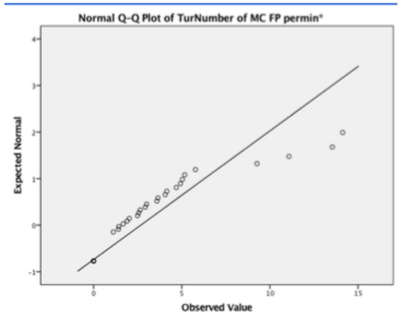


Figure 2. Q-Q plot for English number of MCFP

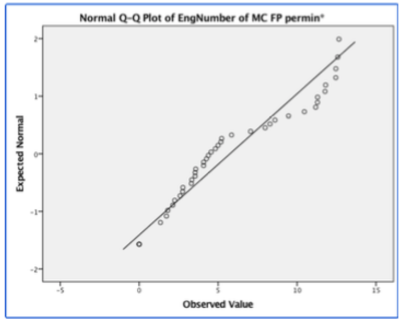


Figure 3. Q-Q plot for Turkish Mean Length of MCFP

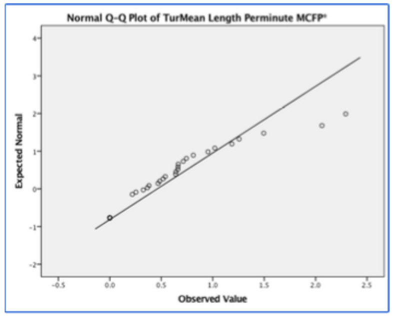


Figure 4. Q-Q plot for English Mean length of MCFP

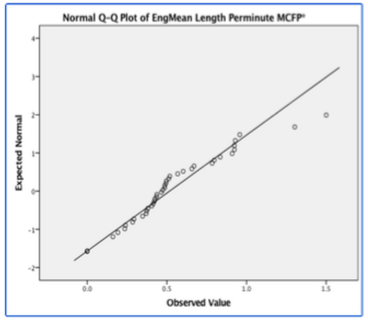


Figure 5. *Q-Q plot for Turkish number of ECFP*

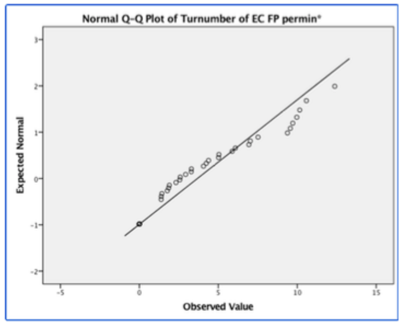


Figure 6. Q-Q plot for English number of ECFP

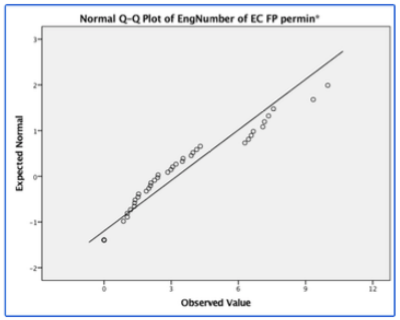


Figure 7. Q-Q plot for Turkish mean length of ECFP

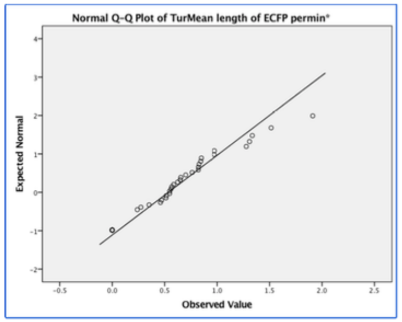


Figure 8. Q-Q plot for English mean length of ECFP

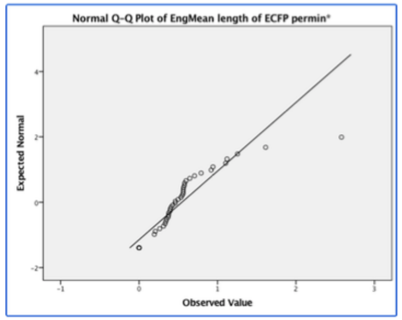


Figure 9. Q-Q plot for Turkish number of MCSP

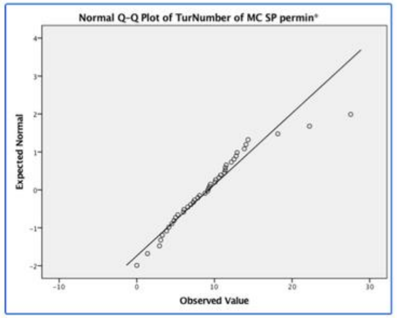


Figure 10. Q-Q plot for English number of MCSP

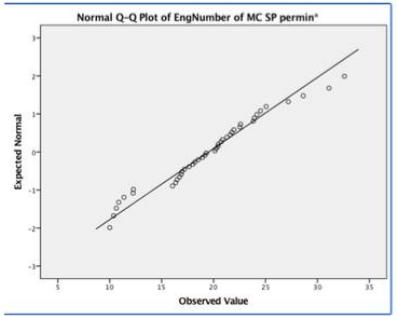


Figure 11. Q-Q plot for Turkish mean length of MCSP

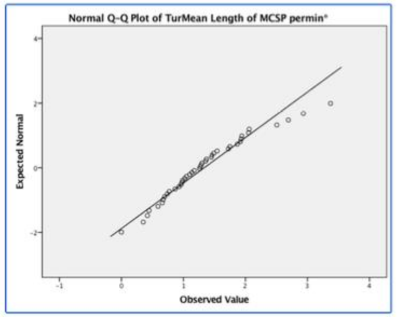


Figure 12. Q-Q plot for English mean length of MCSP

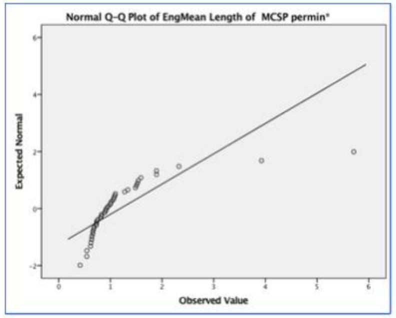


Figure 13. Q-Q plot for Turkish number of ECSP

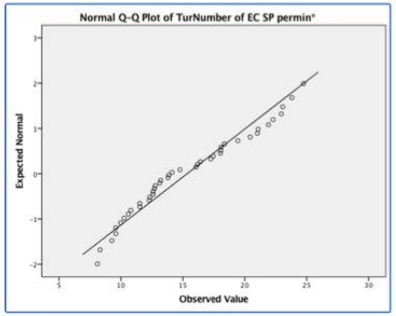


Figure 14. Q-Q plot for English number of ECSP

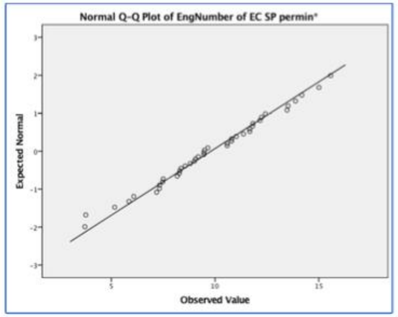


Figure 15. Q-Q plot for Turkish mean length of ECSP

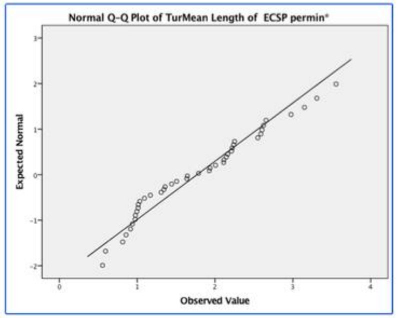


Figure 16. Q-Q plot for English length of EXSP

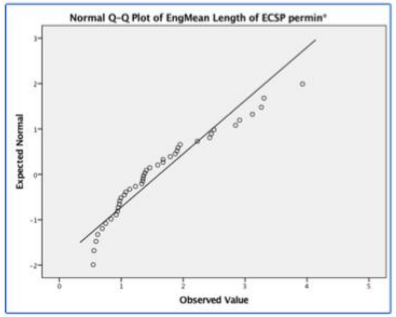


Figure 17. Q-Q plot for Turkish repetitions

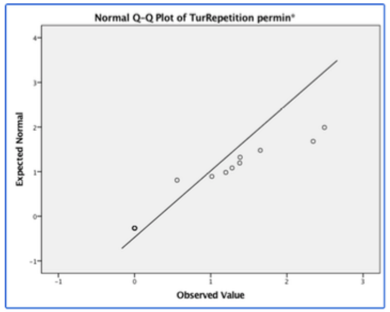


Figure 18. Q-Q plot for English repetitions

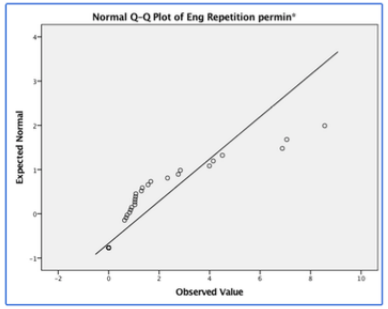


Figure 19. Q-Q plot for Turkish reformulations

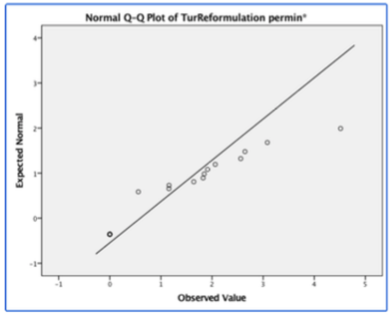


Figure 20. Q-Q plot for English reformulations

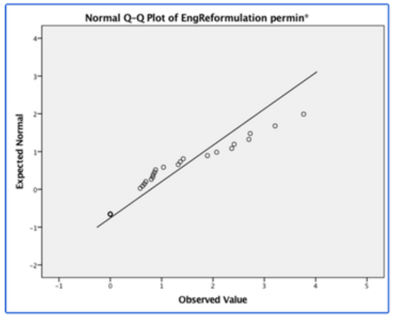


Figure 21. Q-Q plot for Turkish replacements

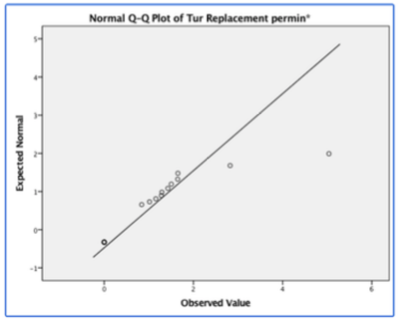


Figure 22. Q-Q plot for English replacements

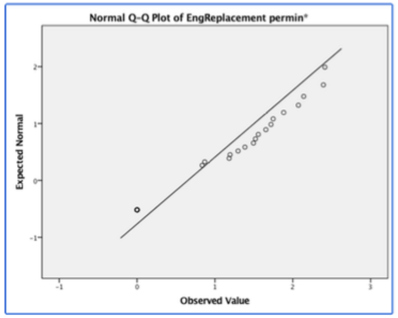


Figure 23. Q-Q plot for Turkish false starts

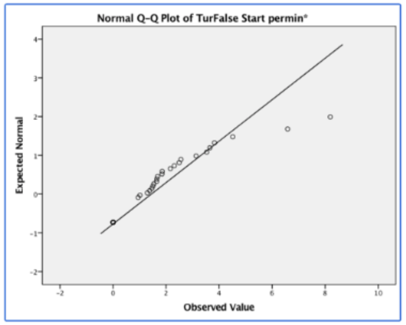


Figure 24. Q-Q plot for English false starts

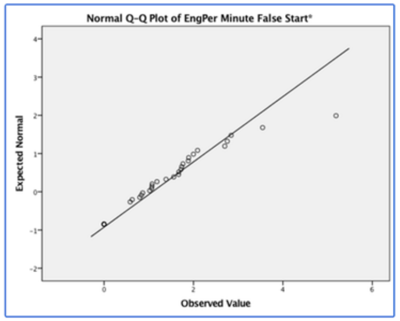


Figure 25. Q-Q plot for Turkish total repair

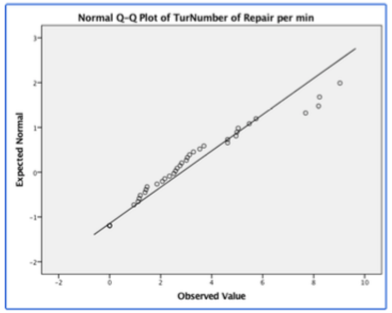


Figure 26. Q-Q plot for English total repair

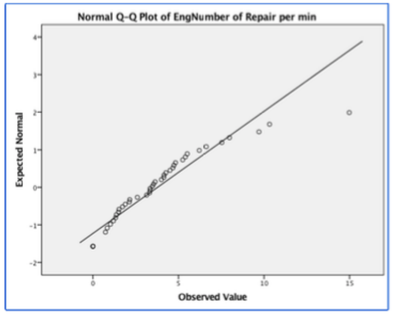


Figure 27. Q-Q plot for Turkish speech rate

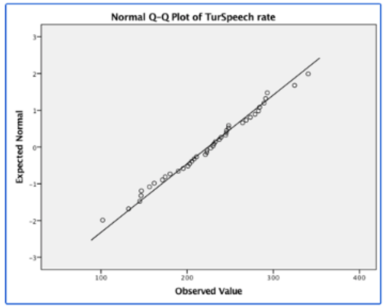


Figure 28. Q-Q plot for English speech rate

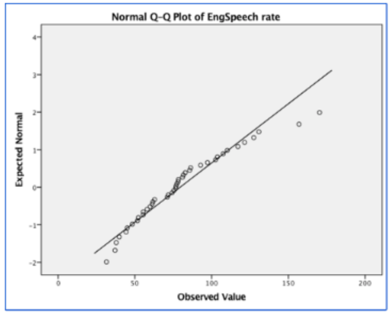


Figure 29. Q-Q plot for Turkish articulation rate

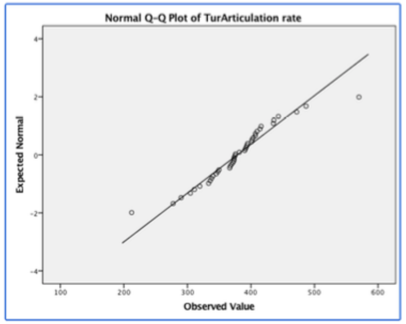


Figure 30. . Q-Q plot for English articulation rate

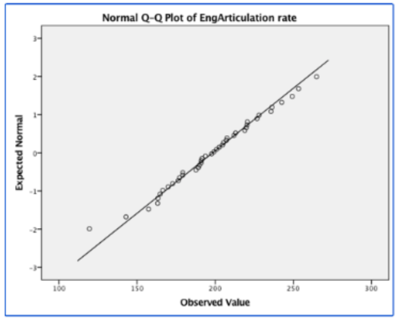


Figure 31. . Q-Q plot for Turkish phonation-time ratio

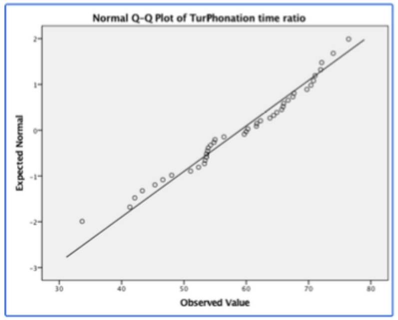


Figure 32. . Q-Q plot for English phonation-time ratio

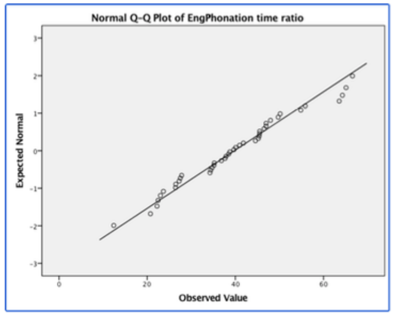


Figure 33. . Q-Q plot for Turkish mean length of runs

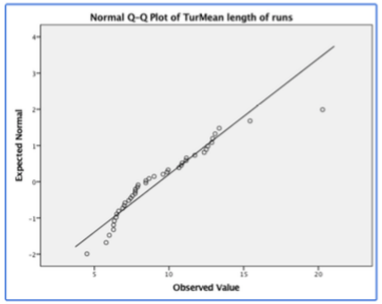


Figure 34. . Q-Q plot for English mean length of runs

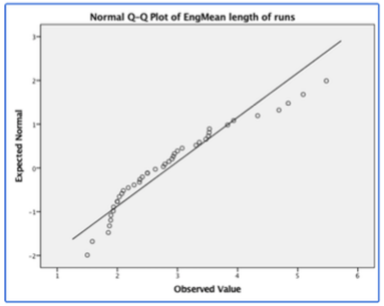


Figure 35. . Q-Q plot for EIT scores

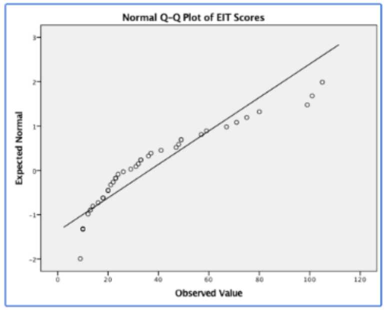


Figure 36. . Q-Q plot for OPT scores

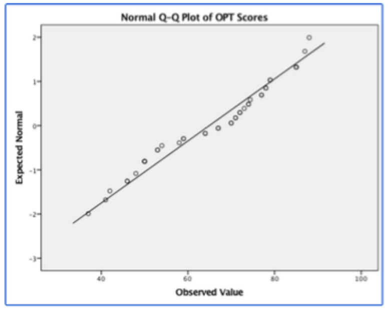
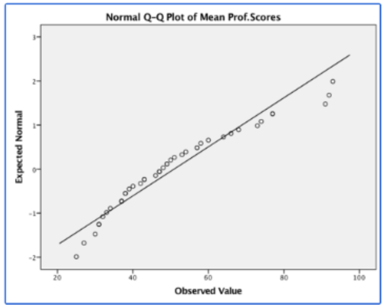


Figure 37. Q-Q plot for mean proficiency scores



Appendix J. Scatterplots obtained for the dependent variables (i.e fluency measures in Turkish and English) of Study 1

Figure 1. Scatterplot for number of MCFP

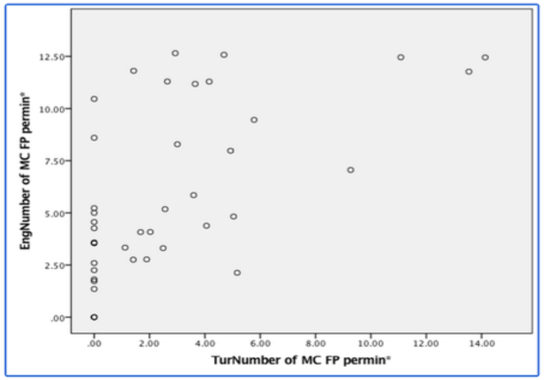


Figure 2. Scatterplot for mean length of MCFP

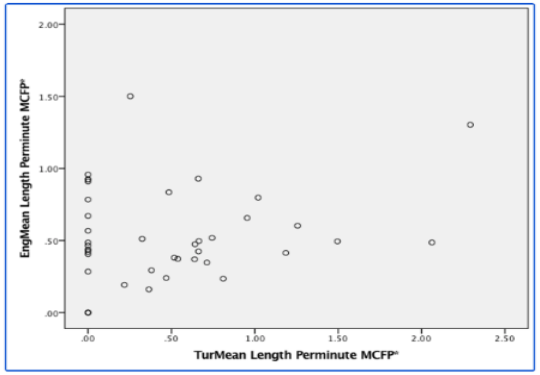


Figure 3. Scatterplot for number of ECFP

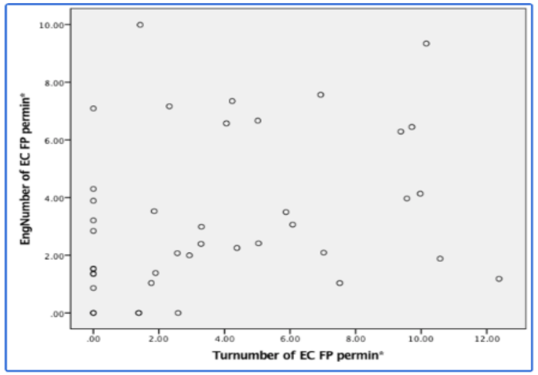


Figure 4. Scatterplot for mean length of ECFP

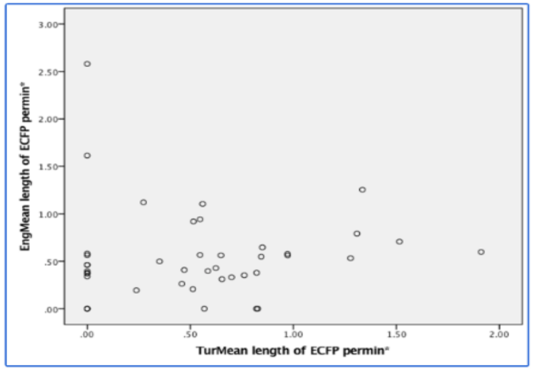


Figure 5. Scatterplot for number of MCSP

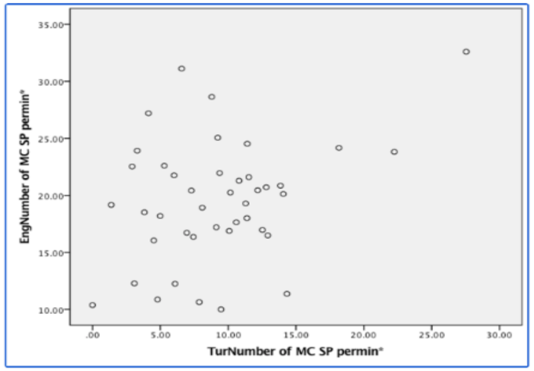


Figure 6. Scatterplot for mean length of MCSP

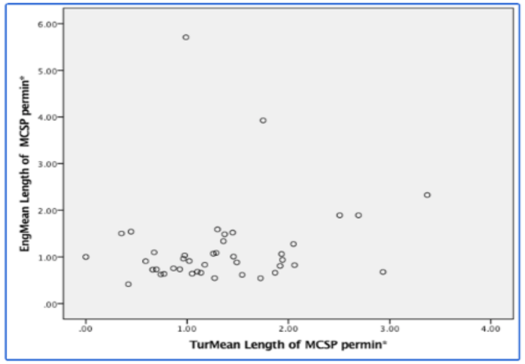


Figure 7. Scatterplot for number of ECSP

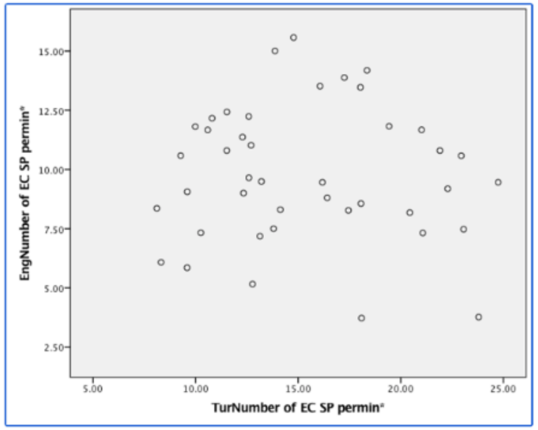


Figure 8. Scatterplot for mean length of ECSP

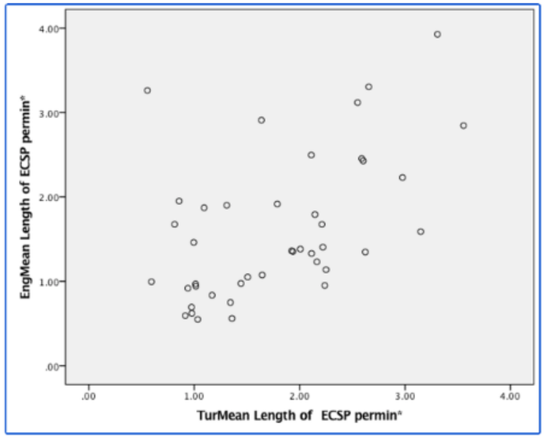


Figure 9. Scatterplot for repetitions

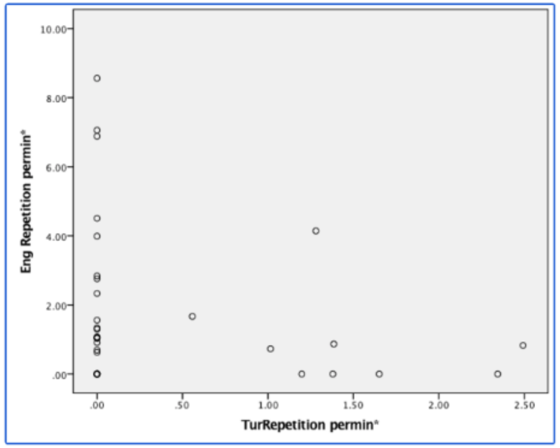


Figure 10. Scatterplot for reformulations

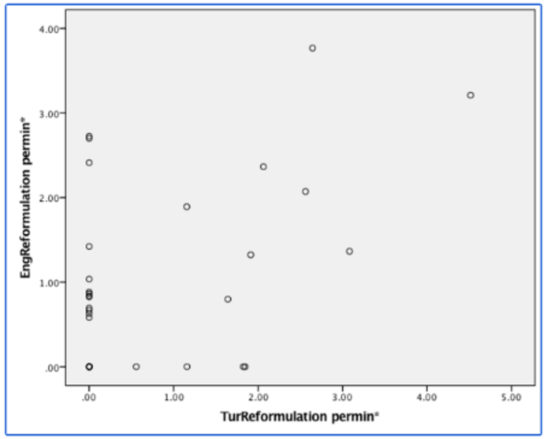


Figure 11. Scatterplot for replacements

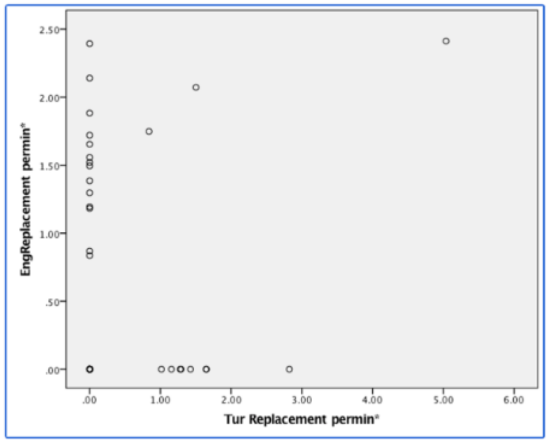


Figure 12. Scatterplot for false starts

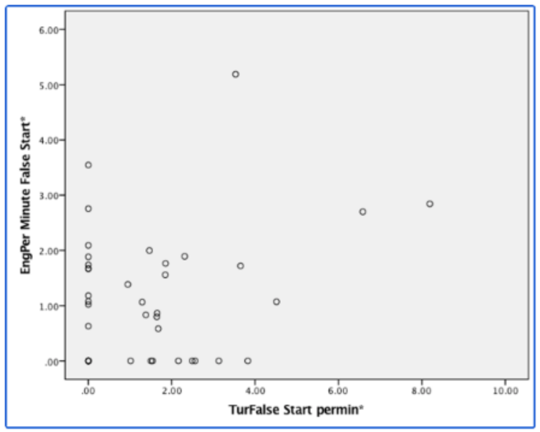


Figure 13. Scatterplot for total repair

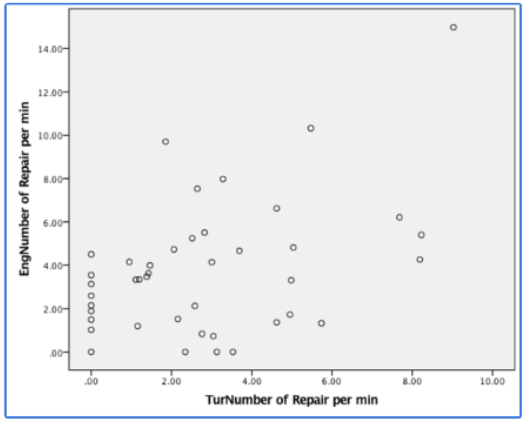


Figure 14. Scatterplot for speech rate

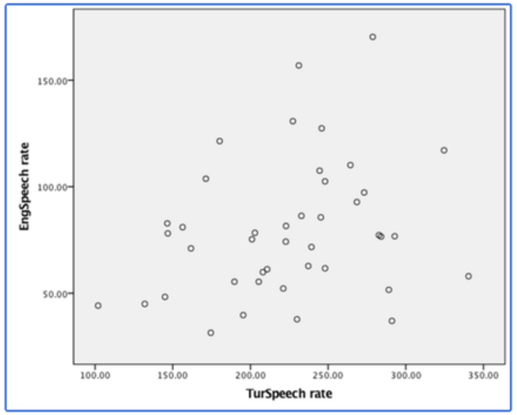


Figure 15. Scatterplot for articulation rate

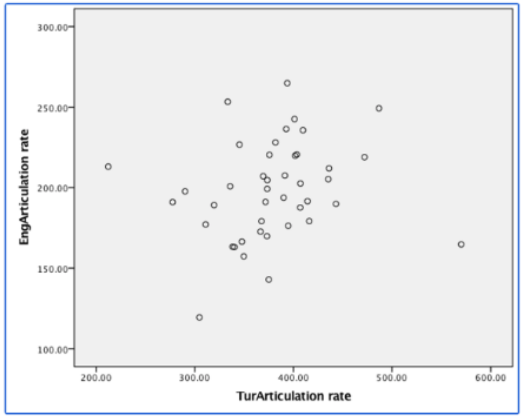


Figure 16. Scatterplot for phonation-time ratio

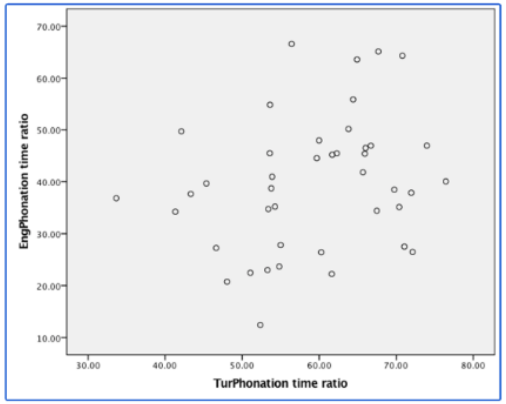
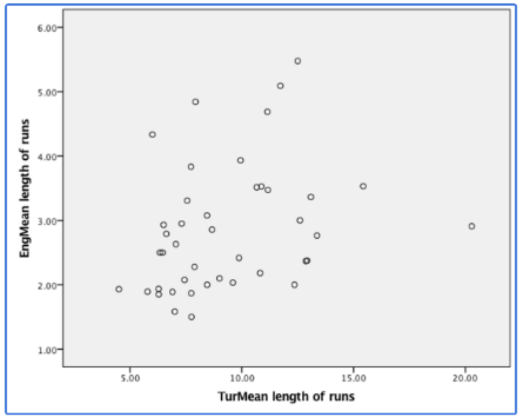


Figure 17. Scatterplot for mean length of runs



Appendix K. Background questionnaire used in Study 2.

School of Literature and Languages
Department of English Language and Applied Linguistics

University of Reading

BACKGROUND QUESTIONNAIRE

Dear Participant,

The following questionnaire is part of my PhD research Project in Applied Linguistics which investigates relationship between first and second language speaking ability. For my Project, I need some information about your language background; so I kindly ask you to complete the questionnaire below.

Please read and tick (X) the box.

☐ *"By completing and returning this questionnaire I understand that I am giving consent for my responses to be used for the purposes of this research project"*

Part. A. General Information

1. Age: ....
2. How long have you been living in the UK (how many months)?

   …………………………………………………………………………………..
3. Level of Education (your most recent educational level/ the latest degree you have finished)

   …………………………………………………………………………….....
4. If you are studying in the UK, please write the degree and year of study. (e.g. 1st year, PhD in …)

   ………………………………………………………………………………………..
5. Known languages other than English (Please, also indicate your level in this language, if any):

   ………………………………………………..

1

Part B. Use of English

1. Age of first exposure to English (Please tick the appropriate age group below).

   ☐ 0-6 ☐ 7-15 ☐ 16-24+

2. Context of first exposure. (Please tick the appropriate option below. You can tick more than one.)

   ☐ at home ☐ online

   ☐ with friends ☐ after arriving in an English-speaking country

   ☐ at school ☐ at work

   ☐ other (please indicate): ………………………………

3. How did you learn English up to this point? You can tick more than one.

   ☐ mainly through formal instruction

   ☐ mainly through interacting with people

   ☐ a mixture of both

   ☐ other (please indicate) …………………

4. Number of years of English instruction that you received before coming to the UK:

   ……………………………………………………………………………………………...

5. How often do you use Turkish and English? Give an estimate in percentage per day (e.g. %60). If any of these does not apply to you (i.e. N/A), put a tick where appropriate.

	at home (with family members)		at university		at work		socializing		social media		TV/radio		daily activities (e.g. shopping, GP, etc.)	
	Turkish	English	Turkish	English	Turkish	English	Turkish	English	Turkish	English	Turkish	English	Turkish	English
%														
N/A														

Additional comments (if have additional questions that you feel should be included here, you write them below with your answers.)

…………………………………………………………………………………………………

THANK YOU FOR YOUR TIME AND CONTRIBUTION

Appendix L. Questionnaire given to L2 teachers and researchers (Study 2).

QUESTIONS

A teacher is going to choose two of three tasks below to give to her students who are at **intermediate level** of English language proficiency. She wants to elicit speech samples from her students to examine the *syntactic structures*, *vocabulary* and the *range of language expressions* they use. Please try to help the teacher by answering the questions below.

The instructions she is going to give the students are:

'*The pictures you see below belong to a Tom and Jerry story. First you have 1 minute to look at these and try to understand what is happening in each picture and in the story. Then, tell the story to your partner, who cannot see the story, in a way that she/he understands what is happening.'*

1) Are these three tasks comparable, in terms of difficulty of understanding the story, from a language teaching perspective? Why?

2) Are these three tasks comparable, in terms of difficulty of performing the task and its linguistic demands (e.g. structures, vocabulary etc.)? Why?

3) Which two tasks can elicit comparable language in terms of vocabulary, syntactic structures, range of language expressions, etc.?

4) Please rate each task in terms of the following aspects for ***intermediate level young adult learners (18-22)*** ofEnglish. Put an X in the relevant columns. (1= very easy/interesting and 6= very difficult/boring)

Task A- Quiet Please

	1	2	3	4	5	6	
easy to understand							**difficult to understand**
easy to tell							**difficult to tell**
interesting to young adult learners							**boring to young adult learners**

TASK B- Best friend

	1	2	3	4	5	6	
easy to understand							**difficult to understand**
easy to tell							**difficult to tell**
interesting to young adult learners							**boring to young adult learners**

TASK C- Jerry and the Goldfish

	1	2	3	4	5	6	
easy to understand							**difficult to understand**
easy to tell							**difficult to tell**
interesting to young adult learners							**boring to young adult learners**

Any other comments:

Thank you very much for your time and cooperation.

TASK A-Quiet Please!

The pictures you see below belong to a Tom and Jerry story. First you have 1 minute to look at these and try to understand what is happening in each picture and in the story. Then, tell the story to your partner, who cannot see the story, in a way that she/he understands what is happening.

The characters: Tom (the cat), Jerry (the mouse) and the Dog

TASK B-Best friend

The pictures you see below belong to a Tom and Jerry story. First you have 1 minute to look at these and try to understand what is happening in each picture and in the story. Then, tell the story to your partner, who cannot see the story, in a way that she/he understands what is happening.

The characters: Tom (the cat), Jerry (the mouse) and the D

TASK C-Jerry and the Goldfish

The pictures you see below belong to a Tom and Jerry story. First you have 1 minute to look at these and try to understand what is happening in each picture and in the story. Then, tell the story to your partner, who cannot see the story, in a way that she/he understands what is happening.

The characters: Tom (the cat), Jerry (the mouse) and the Goldfish

Appendix M. Oral narrative tasks used in Study 2 (Tom and Jerry cartoon scripts).

TASK A-Quiet Please!

The pictures you see below belong to a Tom and Jerry story. First you have 1 minute to look at these and try to understand what is happening in each picture and in the story. Then, tell the story to your partner, who cannot see the story, in a way that she/he understands what is happening.

The characters: Tom (the cat), Jerry (the mouse) and the Dog

TASK B-Best friend

The pictures you see below belong to a Tom and Jerry story. First you have 1 minute to look at these and try to understand what is happening in each picture and in the story. Then, tell the story to your partner, who cannot see the story, in a way that she/he understands what is happening.

The characters: Tom (the cat), Jerry (the mouse) and the D

Appendix N. Backward Digit Span Test (adopted from Awwad, 2017)

Backward Digits Span Test (English Language)

This auditory task is designed to test learners' complex working memory capacity (storage and processing). Participants are required to listen to sets of increased digits and repeat them backwards. Numbers are recorded at one digit per second. Each learners' working memory span is determined based on the last digits set he/she has repeated successfully twice.

Instructions:

You are going to listen to different sets of numbers. I will say the numbers and you have to repeat each set **backwards**. We will start with sets of three digits, and the digits will be increased in sets sizes. When you have two successful attempts, you move to the next set (4 digits), and so on. The test finishes when you fail twice to repeat any of the sets.

For example:

When I say: **"4 5 6"**

You say: **"6 5 4"**

Let me know when you are ready.

Span	**First trial**	**√ / ×**	**Second trial**	**√ / ×**	**Third trial**	**√ / ×**
Three	5. 8. 2.		3. 9. 5.		6. 2. 7.	
Four	3. 9. 1. 5.		4. 8. 2. 6.		1. 9. 7. 3.	
Five	6. 8. 4. 7. 1.		7. 3. 1. 6. 9.		2. 5. 1. 8. 4.	
Six	5. 9. 2. 8. 3. 4.		4. 6. 9. 1. 7. 2.		3. 5. 8. 2. 6. 1.	
Seven	7. 4. 5. 2. 8. 4. 6.		8. 3. 6. 1. 9. 5. 7.		5. 9. 1. 3. 7. 2. 8.	
Eight	9. 2. 5. 1. 8. 7. 5. 3.		1. 6. 8. 2. 9. 3. 7. 4.		8. 1. 4. 9. 2. 5. 7. 3.	
Nine	4. 8. 3. 2. 6. 1. 7. 5. 9.		6. 9. 2. 7. 4. 8. 3. 1. 5.		7. 5. 1. 9. 3. 6. 8. 4. 5.	
Student name						
L2 Backward Digits Span result						

Appendix O. Serial Non-Word Recognition Task (adopted from O'Brien et al., 2006).

Serial Non-Word Recognition Task

You are going to listen to different lists of non-words (words which are not meaningful or not existing). You will hear a list twice, one after another with a short interval between them. When a set is repeated second time, please tell me whether the words that you heard in both lists were in the same order or not. You will start with 4 sets as practice and move on to the actual test.

Let me know when you are ready.

Four non-words (PRACTICE)	P1 (dif)	Bordge tig narp gock Bordge narp tig gock
	P2 (sa)	Keb dern puch ged Keb dern puch ged
	P3 (sa)	Mot chen ped kig Mot chen ped kig
	P4 (dif)	Merd garp tam pib Merd tam garp pib
Actual Test		
Five non-words	P1 (dif)	Terch cam mitch bon derp Terch cam bon mitch derp
	P2(dif)	Choom kerp lork nug gan Choom kerp nug lork gan
	P3 (sa)	Pim tidge gab bock chell Pim tidge gap bock chell

	P4 (dif)	Goot bick mern nuck tep Goot mern bick nuck tep
	P5 (sa)	Loog coll dodge kom meb Loog coll dodge kom meb
	P6 (sa)	Keech chon mup targ bick Keech chon mup targ bick
	P7 (dif)	Coom tord modge gop gick Coom tord gop modge gick
	P8 (sa)	Lod tudge jick norb gaim Lod tudge jick norb gaim
Six non-words	P1 (sa)	Bordge chad nig dack kerp larm Bordge chad nig dack kerp larm
	P2 (dif)	Terdge joop leck norg cham jat Terdge joop norg leck cham jat
	P3 (dif)	Parn meb dorge nug cheem jit Parn dorge meb nug cheem jit
	P4 (sa)	Chon boodge tud lig peb dort Chon boodge tud lig peb dort
	P5 (sa)	Tudge mup chen gerb noog dit Tudge mup chen gerb noog dit
	P6 (sa)	Torm pag jeck derb coll bup Torm pag jeck derb coll bup
	P7 (dif)	Korp teeb nool chorg pim gadge

		Korp teeb nool pim chorg gadge
	P8 (dif)	Padge narp mon chud gop ged Padge narp chud mon gop ged
Seven non-words	P1 (sa)	Toock chad lidge chorg darp nart gub Toock chad lidge chorg darp nart gub
	P2 (dif)	Cark norg mord cham bool lub tep Cark norg mord cham lub bool tep
	P3 (sa)	Charn nig kom jeel gadge lerb mun Charn nig kom jeel gadge lerb mun
	P4 (dif)	Tob gan darch chool juck nord pem Tob gan chool darch juck nord pem
	P5 (dif)	Gell nerg lud pock mitch doob jat Gell nerg lud mitch pock doob jat
	P6 (dif)	Leem cug chordge jert ked darp gock Leem cug chordge ked jert darp gock
	P7 (sa)	Chig nam peb gop jooch lart teed Chig nam peb gop jooch lart teed
	P8 (sa)	Jarm neb gerb chorg mal tudge lon Jarm neb gerb chorg mal tudge lon

Appendix P. Information Sheet given to the participants in Study 2.

School of Literature and Languages
Department of English Language and Applied Linguistics

University of Reading

Researcher:
Zeynep Duran Karaoz
Phone: [redacted]
Email: Zeynep.DuranKaraoz@pgr.reading.ac.uk

Supervisor:
Parvaneh Tavakoli
Phone: [redacted]
Email: p.tavakoli@reading.ac.uk

Department of English Language and Applied Linguistics

URS Building
The University of Reading
Whiteknights, PO Box 219
Reading RG6 6AW

Phone [redacted]

[redacted] -[redacted]

Email appling@reading.ac.uk
p.a.thompson@reading.ac.uk

INFORMATION SHEET

This project aims to understand if there is a relationship between first and second language speaking ability. You have been selected to take part in this study because you speak Turkish as your native language and English as your second language. 60 (approx.) learners, aged between 22-35, will participate in the study. Data collection procedure will be run on two separate days. On Day 1, you will be given a short questionnaire and two English language proficiency tests; one after another with short breaks in between. In one of these tests, you will be asked to answer a set of questions on paper. In the second one, you will hear a set of sentences and will repeat them after hearing each sentence. Your speech will be audio-recorded to be scored. You will be contacted to meet on a different day which is convenient for you.

On Day 2, you will be met individually on a one-to-one setting, and will be asked to look at a set of pictures and tell two short stories; with one being in Turkish and the other in English. Then, you will be asked to take two memory tests, where you will listen to a set of numbers in one of them and a set of non-words in the other; in the first one, you need to repeat the numbers in the reverse order while in the second one, you will tell the researcher if the sets are in the same order or not. In story telling tasks and the memory tests in the second phase of the study, your performance will be audio-recorded to be analysed.

All the data collected will be securely kept in a password-protected computer, and will be accessed only by the investigator, Zeynep Duran Karaoz, and her supervisors. The data will be used for research purposes only and will be destroyed immediately after the completion of the thesis (in five years' time). Your names will not be mentioned in the project and you will be able to withdraw from the study at any time you wish to. Your privacy and confidentiality will be carefully observed.

This project has been subject to ethical review by the School Ethics and Research Committee, and has been allowed to proceed under the exceptions procedure as outlined in paragraph 6 of the University's *Notes for Guidance* on research ethics.

If you have any queries or wish to clarify anything about the study, please feel free to contact my supervisor at the address above or by email at [include supervisor's email address here]

Signed