UNIVERSITAT DE BARCELONA

Department of Modern Languages and Literatures and

English Studies

M.A. Thesis

The impact of complexity, accuracy and fluency (CAF) on comprehensibility and perceived fluency in the case of L2-Greek: a partial replication study

Author: Katerina Defto

Supervisors: Dr. Roger Gilabert Guerrero & Dra. Maria Andria

UNIVERSITAT DE BARCELONA

Facultat de Filologia
Departament Filologia Anglesa i Alemanya

Gran Via de les Corts Catalanes, 585
08007 Barcelona, SPAIN
Tel. +34 934 035 686
Fax +34 933 171 249

Màster Oficial en Lingüística Aplicada
i Adquisició de Llengües en Contextos Multilingües
LAALCM

Roger Gilabert i Maria Andria, com a supervisors/es del treball (Tesina de Màster) presentat com a requeriment per a l'avaluació de l'assignatura **Projecte de Recerca en Lingüística Aplicada**

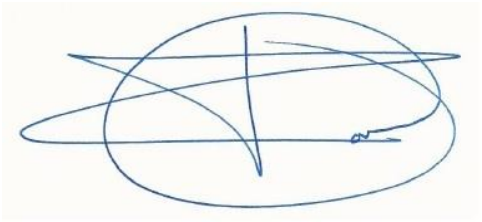
presentat per l'alumne/a: **Katerina Defto**

amb el títol de: *The impact of complexity, accuracy and fluency (CAF) on comprehensibility and perceived fluency in the case of L2-Greek: a partial replication study*

certifiquem que hem llegit el treball i l'aprovem perquè pugui ser presentat per a la seva defensa pública al juliol de 2020.

I perquè consti i tingui els efectes oportuns signem aquest certificat en

Barcelona, a 1 de juliol de 2020



Dr. Roger Gilabert Guerrero **Dra. Maria Andria**

UNIVERSITAT DE BARCELONA

Facultat de Filologia i Comunicació
Dept. Llengües i Lit. Modernes i Estudis Anglesos

Gran Via de les Corts Catalanes, 585
08007 Barcelona, SPAIN
Tel. +34 934 035 686
Fax +34 933 171 249

Official MA programme in

Applied Linguistics and Language Acquisition in Multilingual Contexts (LAALCM)

Universitat de Barcelona

Non-Plagiarism Statement

This form must be completed, dated and signed and must be included at the beginning of every copy of the MA Thesis you submit for assessment.

Name and surnames:	Katerina Defto
MA Thesis title:	The impact of complexity, accuracy and fluency (CAF) on comprehensibility and perceived fluency in the case of L2-Greek: a partial replication study
Supervisor:	Dr. Roger Gilabert Guerrero & Dra. Maria Andria

I HEREBY DECLARE THAT:

- This MA Thesis that I am submitting for assessment is entirely my own work and I have written it completely by myself.

- I have not previously submitted this work or any version of it for assessment in any other programme or institution.

- I have not used any other sources or resources than the ones mentioned.

- I have identified and included the source of all facts, ideas, opinions and viewpoints of others through in-text referencing and the relevant sources are all included in the list of references at the end of my work. Direct quotations from books, journal articles, internet sources or any other source whatsoever are acknowledged and the sources cited are identified in the list of references.

I understand that plagiarism and copying are serious offences. In case of proof that this MA Thesis fails to comply with this declaration, either as negligence or as a deliberate act, I understand that the examiner has the right to exclude me from the assessment act and consequently all research activities conducted for this course will be declared null and the MA Thesis will not be presented for public defense, thus obtaining the lowest qualification

Date: 1/07/2020 **Signature:**

Acknowledgments

This research has been carried out in the framework of the LETEGR2 project (*Learning, Teaching, and Learning to Teach in Greek as a Second/Foreign Language: Evidence from different learning contexts)* and it has been financially supported by the General Secretariat for Research and Technology (GSRT) and the Hellenic Foundation for Research and Innovation (HFRI) (Research Code: 1656).

First and foremost, I would like to express my gratitude towards my supervisor, Dr. Roger Gilabert Guerrero, for all the support and advice I have received from him this last year. I would have never made it to the end of this project, had it not been for his guidance. I also deeply thank my supervisor, Dra. Maria Andria, for providing us with the data from the LETEGR2 project, that made it possible for this study to be realised, and for helping and supporting me every step of the way. I am also deeply grateful to Dr. Juan Carlos Mora for being always available to share his knowledge and expertise. His comments and recommendations have been of great value to me. I would also like to sincerely thank Panagiotis Panagopoulos, member of the research project LETEGR2, who assisted me incredibly with the coding of complexity and accuracy. A special mention to my friend and classmate Paola Mannarelli who has always been there to talk over any little doubt and concern I had and for literally saving my life more than once, despite her overloaded schedule. Last but not least, I need to express my very profound gratitude towards my family and especially my mother. No words could describe how much I owe her.

Abstract

The present Master thesis aimed to partially replicate the article by Suzuki and Kormos (2019) on the linguistic dimensions of comprehensibility and perceived fluency. The distinguishability among the two constructs as well as their associations to complexity, accuracy and fluency (CAF) were investigated in the case of L2-Greek picture/descriptive speech. Speech stimuli from 68 Spanish/Catalan L2-Greek learners was presented to 8 naïve native judges to be evaluated with regards to comprehensibility and perceived fluency in a 9-point scale and was objectively analysed in terms of CAF measurements. Correlation analysis showed that most of the CAF variables are more or less correlated with both comprehensibility and perceived fluency and confirmed a strong association among the two constructs. However, judges were stricter when judging fluency than when judging comprehensibility. Furthermore, a series of multiple regression analyses revealed that within-clause pause ratio, grammatical accuracy and lexical complexity are the strongest predictors of comprehensibility, while grammatical accuracy, within-clause pause ratio, lexical complexity and lexical error rate best predict perceived fluency.

Keywords: Comprehensibility judgments, perceived fluency judgments, CAF measurements, naïve native judges, correlation, regression analysis

Table of Contents

I. Introduction

Fluency and comprehensibility constitute two constructs of great importance in the field of second language acquisition. Not only do they provide researchers with an insight into the specific features of L2 oral speech, but they also help teachers set a well-established and realistic goal for learners in an L2 teaching context (Derwing & Munro, 2013; Isaacs & Trofimovich, 2012; Saito et al, 2018). Since pursuing an overall native-like competence is highly ambitious, especially for late learners, and only few talented people may succeed (Abrahamsson & Hyltenstam, 2008; DeKeyser, 2000; Flege et al., 1995;), it is reasonable to focus our attention on how to become more fluent and easily intelligible in an L2 (Derwing & Munro, 2009; Levis, 2005). It is noteworthy that both constructs are used as descriptors in many language tests of widely recognized status (e.g. IELTS, TOEFL iBT) and could have an important impact on the L2 speakers' career opportunities (Saito, Trofimovich & Isaacs, 2016).

What makes speech be perceived as fluent and comprehensible though? The linguistic correlations of comprehensibility and perceived fluency vary in the literature due to many methodological factors, such as the nature of the speech elicitation tasks, the length of the speech stimuli and the operationalization of the two concepts across studies (Suzuki & Kormos, 2019). In addition to these, the lack of studies examining listeners' perception of L2 oral speech in a variety of languages is an unfortunate and yet inevitable fact, that impedes even more our in-depth understanding of the issue. With regards to fluency, for example, research indicates that listeners' judgments are influenced by different acoustic features across languages based on the unique and specific speech delivery patterns of each language code (Préfontaine et al., 2016). Following this line of research, which calls for more studies on comprehensibility and perceived fluency of multiple language sets, it was decided to explore what happens in

the case of L2-Greek, a relatively under-researched language in the field. In particular, in the present study we investigate the linguistic correlates in terms of complexity, accuracy and fluency -henceforth CAF- of comprehensibility and perceived fluency in the case of Spanish/Catalan learners' L2-Greek speech. To our knowledge no studies have been conducted so far on the perception of L2-Greek oral speech and it is our aim to provide data that come from a language system that has never been investigated to this regard. Moreover, it is our objective our findings to be comparable to a large extent with the previous conducted research, hence we attempted to replicate the study by Suzuki and Kormos (2019). A subjective-objective approach is adopted by which global rating of speaking performances and objective measurements of the linguistic features of these same performances are brought together (De Jong et al., 2012, p.6). On the other hand, a subjective-subjective approach includes only subjective ratings by the judges on both comprehensive concepts of speaking performances and the specific linguistic features of these same performances.

In the literature review that follows we present in two consecutive subsections the most common definitions and research findings of fluency and comprehensibility. A section devoted to CAF follows and another section that refers to the original study by Suzuki and Kormos (2019) concludes the literature review. We next move on to the research questions of this paper and the methodology part, where a description of the participants, the materials and the procedures is given. The linguistic and statistical analysis of the collected data comes next, in which we explain the CAF measurements that were calculated and the followed statistical analysis. We then briefly present the results before finishing with the discussion and the conclusion, while the limitations of the study conclude this whole paper.

II. Literature review

a. Fluency

The notion of fluency, as compared to the one of comprehensibility, is a notion most people are familiar with. The word itself is used by ordinary people on many occasions in order to refer to one's oral language skills in an L2. Therefore, there seems to be a consensus in popular belief that fluency accounts for overall oral proficiency (Chambers, 1997). Research in the field has shown that even professional language teachers view fluency in this broader sense and frequently use the term of fluency as an alternative to their students' general speaking ability (Tavakoli & Hunter, 2018). However, such an approach beclouds many of the specific features that characterize comprehensively the concept of fluency. It is hard to find in the literature a single definition that includes and briefly describes all that fluency is. Instead each scholar points out and adds a different aspect of this multifaceted construct.

According to Lennon (1990) fluency can be defined either in a broader, more general sense or more narrowly. In the broad sense fluency is equated to overall proficiency referring to the grammar, vocabulary and accent skills of the speaker. This notion of fluency usually coincides with what most non-specialists think it to be, as it has already been mentioned above. In the narrow sense, though, fluency is viewed as only one component of one's oral skills, which highly depends on the temporal features of the speech, such as speed, filled and unfilled pauses, repetitions or self-corrections among others. It is a 'performance phenomenon' that reflects the listener's impressions on how easily and smoothly the speaker's speech planning and speech production mechanisms are functioning (Lennon, 1990). Moreover, most definitions of fluency have repetitively highlighted qualitative properties of speech such as smoothness, flow, effortlessness

and automaticity as being essential key elements to the concept (Chambers, 1997; Crystal, 1987; Lennon, 1990; Lennon, 2000; Schmidt, 1992;).

By adopting a cognitive-based perspective three distinct levels of fluency are defined: cognitive, utterance and perceived fluency (Segalowitz, 2010). Cognitive fluency refers to "the efficiency of operation of the underlying processes responsible for the production of utterances". Utterance fluency is defined as "the features of utterances that reflect the speaker´s cognitive fluency'. Lastly, perceived fluency refers to "the interferences made by listeners about speakers' cognitive fluency based on their perceptions of their utterance fluency" (Segalowitz, 2010, p.165). Furthermore, utterance fluency is further divided into: i. Breakdown fluency, that is related to the amount and quality characteristics of filled and unfilled pauses encountered in a speech signal, ii. Speed fluency, that is basically how fast the speech is delivered, and iii. Repair fluency, that is concerned with the number of dysfluencies (self-corrections, repetitions etc.) found in speech (Tavakoli & Skehan, 2005).

Studies in the field of perceived fluency research have so far widely and unanimously attribute high significance to speech rate and other speed fluency measures, as strong predictors of fluency judgment scores (Bosker et al., 2013, Derwing, Rossiter et al., 2004; Kormos & Dénes, 2004; Préfontaine et al., 2016; Saito et al., 2018; Suzuki & Kormos, 2019). Similarly, there seems to be a general consensus in the literature with regards to repetitions, self-corrections and other dysfluency phenomena, whose importance on perceived fluency evaluations appears to be limited (Bosker et al., 2013; Derwing, Rossiter et al., 2004; Kormos & Dénes, 2004; Saito et al., 2018; Suzuki & Kormos, 2019). These results are hard to align with the findings that listeners in fact do notice repair fluency behaviour in the part of the speaker and they actually tend to judge it positively (Bosker et al., 2013; Préfontaine & Kormos, 2016).

Nonetheless, the relative weight of other temporal dimensions on subjective fluency ratings is less clear with studies demonstrating remarkable deviations among them or even directly contradicting each other. One of the most debatable aspects of fluency perception is the role of the pauses (breakdown fluency). In the first of the four-experiment study conducted by Bosker et al. (2013) the number and duration of pauses were proved to best predict the subjective fluency scores explaining alone the largest part of the variance (R^2 =0.5917). Similarly, participants' mid-clause pausing behaviour in Suzuki and Kormos' (2019) article strongly correlated with fluency judgments indicating the importance of pause location, when evaluating fluency. The importance of pause location is additionally found to change with regards to the speakers' proficiency level. Research findings support that final-clause pauses differentiate between beginner- and intermediate-level speakers, while mid-clause pauses differentiate intermediate to advance L2-speech (Saito et al., 2018).

On the other hand, a number of studies find no correlation or mixed results between fluency scores and breakdown fluency measures. Teachers' fluency perception of 16 Hungarian L2-English learners' oral performance was not influenced by the number of filled or unfilled pauses according to Kormos and Dénes (2004). However, mean length of pauses was proved to correlate with the composite score of fluency ratings, only because it influenced significantly the assessments of two out of the 6 judges. Derwing, Rossiter et al. (2004) as well found no independent contribution of the pause variable to the fluency judgments of low-proficiency L2-speech by 20 Mandarin speakers of English across three different task types; yet still combined with pruned syllables accounted together for 69% and 68% of the variance in the picture description task and in the monologue task correspondently. The ambiguity of the importance of pauses is also underlined in Préfontaine et al. (2016), where the frequency of the pauses of L2-

French speech was the least important predictor for the dependent variables with mean length of run and articulation rate being the most influential ones. Average pause time, though, was positively correlated with perceived fluency implying a cross-linguistic influence specific to French.

Finally, there is evidence that other linguistic features of oral speech might account in a significant way for the variance of L2 fluency scores. The number of stressed words per minute was discovered to be one of the strongest predictors of fluency judgments in the study by Kormos and Dénes, that was proposed by them as "the new quick and easy way of establishing fluency" (2004:160). Other dimensions that have been found to be important in some cases include target-like rhythm and prosody (Préfontaine & Kormos, 2016), grammatical accuracy (Préfontaine & Kormos, 2016; Suzuki & Kormos, 2019) and lexical diversity (Kormos & Dénes, 2004).

In sum perceived fluency has been in some cases clearly associated with breakdown fluency, while other studies suggest otherwise. Other dimensions of speech such as grammar, lexis, speed fluency, rhythm and prosody have also impacted raters' scores of perceived fluency in several studies. As far as we know, repair fluency has not been reported as a strong predictor of perceived fluency. However, further investigation of dysfluency phenomena is demanded based on findings from qualitative research, which indicated that they are not going unnoticed.

b. Comprehensibility

As compared to the concept of fluency, comprehensibility is a more straightforward concept to define. Nonetheless, it is closely related to intelligibility and sometimes there is a confusion between the two terms. In a broad "non-technical" sense intelligibility refers to whether or not a listener can understand a speaker, a definition that can hardly distinguish it from comprehensibility (Levis, 2006). It is, however, quite clear where

the difference between the two lies once we become more specific. While intelligibility, in the narrow sense, is concerned with the amount of speech a listener eventually understands, comprehensibility focuses on the amount of difficulty the process of understanding requires in the listener's part. In other words, comprehensibility reflects "the listeners' perception of how easy or difficult it is to understand a given speech sample" (Derwing & Munro, 2009, p.478). Even if one comes to a total understanding of an utterance, it is the amount of time and effort one needed that determines comprehensibility and distinguishes it from intelligibility. Moreover, intelligibility is traditionally assessed with orthographical transcriptions, writing summaries or answering comprehension questions, while comprehensibility studies use scalar ratings (Derwing & Munro, 2009).

In addition, comprehensibility goes hand to hand with the construct of accentedness, which refers to the degree of distance between L2- speech and native-like speech with regards to pronunciation. In most studies comprehensibility and accentedness are examined together with researchers coming to the conclusion that, although accentedness and comprehensibility are closely related, they are two distinct and independent constructs. A heavily accented speech may by all means remain highly comprehensible (Derwing & Munro, 2009; Kang et al., 2010). However, further research needs to be done in this direction, since there are indications in the literature that the relationship between comprehensibility and accentedness is far more complex and specific to task demands (Crowther, Trofimovich, Saito & Isaacs, 2018).

Research in the field of comprehensibility provides us with some interesting points. Most studies in this area point in the direction that native judges -either experienced or not- take into account a plethora of linguistic features spanning from pronunciation, prosody, stress and phonology in general to lexical, morphological and other temporal

phenomena (e.g. Crowther, Trofimovich, Saito & Isaacs, 2018; Saito, Trofimovich, & Isaac, 2016; Suzuki & Kormos, 2019). The specific role of lexical dimensions on comprehensibility, for instance, were sought in detail in the study conducted by Saito, Webb, Trofimovich and Isaacs (2016). By controlling for the effect of phonology aspects they were able to isolate and examine only the effect of lexical factors. The results revealed a strong association of comprehensibility with lexical appropriateness, including lemma choice and morphology of a word, and with the category of sense relations, that is to say the number of different meanings (polysemy) attributed to a word by the speaker.

What is more, like fluency comprehensibility seems to be affected by the speakers' proficiency level with different acoustic and linguistic features being important at different stages. Saito, Trofimovich and Isaacs (2016) examined the L2-English speech by 120 Japanese speakers in a picture description task and resulted that word stress and intonation significantly correlated with comprehensibility scores for all proficiency levels. Moreover, speech rate, lexical appropriateness and prosody were found important for beginner to intermediate levels, while good prosody, segmental precision and grammatical accuracy for intermediate to advanced levels. The importance of word stress at all proficiency levels as well as the distinct influence of vocabulary for low-proficient speakers and grammar for advanced speakers had also been reported in a previous study on Francophones (Isaacs & Trofimovich, 2012). These findings clearly provide evidence that, although pronunciation aspects are always significant for comprehensibility judgments, lexis and grammar are crucial for different stages of language acquisition.

c. CAF

The theoretical framework of CAF has been used repeatedly in the field of SLA to measure L2 proficiency and L2 language development and it has been employed as well on comprehensibility and perceived fluency studies. It was first introduced by Skehan twenty years ago and it consists of three fundamental constructs: complexity; the ability to use a range of sophisticated structures and vocabulary, accuracy; the ability to use target-like and error-free language, and fluency; the ability to produce the L2 with native-like rapidity, pausing, hesitation, or reformulation (Housen, Kuiken & Vedder, 2012, p. 2). Defining CAF dimensions, however, can get a lot more complicated. Each of the dimensions is multi-layered and multifaceted. Complexity is a construct rather hard to define with an elaborated taxonomy. The two main divisions are that of relative or cognitive complexity and that of absolute or linguistic complexity. The former refers to the mental effort a linguistic item requires to be processed and acquired during L2 learning and it is thus a synonym to difficulty, while the later refers to the absolute number of components a language feature or a language system consists of as well as the number of the relations between them (Bulté & Housen, 2012; Housen & Simoens, 2016). Most studies using CAF have traditionally focused on the lower branches of linguistic complexity and measure it based on the distinction between structural and lexical complexity. On the other hand, accuracy is probably more accurate to be defined as appropriateness and acceptability considering that there are different types of deviations or else errors from a target-L2, some of which are more tolerable than others (Housen, Kuiken & Vedder, 2012, p. 4). With regards to fluency, the complexity of its concept has been discussed above. We will only remind here the three subdivisions of utterance fluency because they are the most relevant to the CAF framework, these are: speed fluency, breakdown fluency and repair fluency.

Moreover, CAF components have been found to be affected by a variety of factors spanning from inherited properties of the linguistic items and structures to factors related to the learner characteristics, the teaching methodology, the characteristics of the input and the conditions of the task. In task-based research the mode of the task (monologic/dialogic), the cognitive complexity of the task (e.g. number of items in a narrative) or other factors related to task conditions in general (e.g. preparation time) have been shown to have an impact on CAF measurements (Gilabert 2007; Gilabert, Barón & Levkina, 2011; Robinson, 2005).

Finally, two are the main methodological issues that beset the CAF field in general: i. the way CAF components are measured, and ii. the selection of the CAF measurements in a study. Regarding the measuring of the CAF components, it is sufficing to say that there has been a lot of criticism in regard to their validity and new measurements that are more targeted and specific rather than general and global have been suggested (Lambert & Kormos, 2014; Norris & Ortega, 2009;). Moreover, it is important that researchers carefully select the measurements of CAF they use so that they avoid possible collinearity between them (Bosker et al., 2013; Norris & Ortega, 2009; Suzuki & Kormos, 2019).

d. The original article by Suzuki and Kormos

The current study was motivated by the article by Suzuki and Kormos (2019). It is a partial-replication of the aforementioned article, following calls in the field of second language acquisition for replication studies (*Language Teaching* Review Panel, 2008). On that research project the researchers investigated the linguistic dimensions of comprehensibility and perceived fluency in the context of 40 L2-English Japanese speakers. They analysed with a cluster of objective measures five linguistic dimensions: complexity, accuracy, fluency, pronunciation and discourse features. An argumentative

task was used in which participants had to express their opinions on a given statement. Once the speech data was collected, 10 naïve native speakers of English rated on a 1-9-point scale the performances with regards to comprehensibility and fluency. Stepwise regression analysis revealed that both constructs were associated with grammatical accuracy, breakdown fluency and pronunciation. Moreover, comprehensibility was best predicted by articulation rate, while fluency by the frequency of mid-clause pauses. Secondarily, comprehensibility and fluency judgements were found to strongly correlate with each other and native judges were proved to be significantly more lenient when they assessed comprehensibility in comparison to their behaviour during fluency evaluation.

Having the article above as our guide, a number of modifications and adaptations were decided in order to meet the demands and the purposes of the present study. Two important changes were performed; participants came from a different L1/L2 background and the type of the speech elicitation task was a descriptive/narrative task instead of an argumentative one. In essence we followed the same structure and approach with minor accommodations.

III. Research questions

The research questions of the present article were formed based on the research questions posed on the aforementioned article by Suzuki and Kormos (2019). They were adapted to the population under investigation and the task used on this study and were worded as follows:

1. To what extent are comprehensibility and fluency distinguishable for naïve native listeners in the case of Spanish/Catalan learners' L2-Greek narrative/descriptive speech?
2. How are complexity, accuracy and fluency dimensions of performance associated with comprehensibility of Spanish/Catalan learners' L2-Greek narrative/descriptive speech?
3. How are complexity, accuracy and fluency dimensions of performance associated with perceived fluency of Spanish/Catalan learners' L2-Greek narrative/descriptive speech?

IV. Methodology

a. Participants

L2 speakers

The participants were 68 (37 males and 31 females) Spanish/Catalan bilinguals or L1-Spanish speakers of a variety of ages ranging from early twenty to late seventy. According to the so far coded data (about 55%) participants' mean age is 47 years old with minimum and maximum reported age being 20 and 78 respectively. They were studying Modern Greek as a Foreign Language in a formal context, at two language schools in Barcelona, Spain. The two language schools follow the same proficiency level classification and curriculum. The participants belonged to different proficiency levels. According to Common European Framework of Reference for Languages (Council of Europe, 2001) 41 of the participants were classified as B1 speakers, 20 as B2 and 7 as C1. Their motivation for studying Greek varies a lot but the most commonly reported motives were: i. their love for Greece and Greek culture, ii. the constant

connection they keep with the country by spending their vacations there, iii. family bonds through marriage with Greek people (either theirs or their children's), iv. job related factors (some of them were teachers or students of classical Greek and they were interested in expanding their knowledge). It is noteworthy that participants' most reported motivation for learning Greek as a foreign language is their intrinsic genuine interest in the country and its culture, which might be justified by the participants' quite advanced mean age. Older people free form the anxiety of career success and family responsibilities are more likely to engage in second language learning due to their personal interests and desires (Andria, 2014).

Native judges

A total number of 8 individuals (6 females, 2 males) were recruited to serve as judges for the needs of the present study. All of them belonged to the writer's broader social network with some of them participating after personally being asked, while others after responding to a relevant publication by the author in Facebook. The mean age of the judges is 23.125 years old and their age range fluctuates from 21 to 25 years. All of them were born and raised in Greece with both of their parents being native Greek speakers.

Furthermore, for the purpose of this study it was essential to secure that all participants would appertain to the category of inexperienced, naïve judges. According to the background questionnaire none of the individuals is a Greek teacher or has done any studies relative to education or linguistics. Nonetheless, two of them reported to have taught voluntarily and for a limited period of time Greek to foreigners in the past. Moreover, it was reported no significant knowledge of Spanish or Catalan and no significant degree of familiarity with Spanish-accented Greek. In the question with regards to their language knowledge, in particular, they all selected 1 in a 9-point scale

(1=I don't speak at all, 9=I speak fluently) for Catalan and 1-3 (mean=1.5) for Spanish. With regards to familiarity with Spanish language and Spanish-accented Greek, however, their score is a little bit higher varying from 3 to 7 (mean=4.625) and from 2 to 7 (mean=4.25) correspondently. Additionally, judges' mean score for the amount of daily exposure to Spanish-accented Greek was no more than 1.75 (range= 1-3), however, the mean score for judges' daily exposure to foreign-accented Greek in general reached up to 4.75 (range=2-7). Overall, the mean values reported above hardly ever reach the middle of the 9-point scale, and so we can conclude that judges in general were only minimally familiar with Spanish and Spanish-accented Greek.

b. Materials

Picture narrative

The L2 speech elicitation task was a picture-description narrative known as the dog story. It consists of six serial pictures that narrate the story of a boy and girl who are getting ready for a picnic preparing sandwiches. Once they are ready to go, their mum comes in holding a map and the three of them start talking over the route. That is when their dog gets into the basket unnoticed. When the children reach their destination and want to eat the sandwiches, they realise that the dog had been in the basket all along and had eaten all the food (Heaton, 1966 op cit Muñoz 2006).

Speech stimuli

The speech data of 85 L2-Greek speakers narrating the dog story was eventually received in the form of MP3 files. All data was converted to WAV files using Audacity. A small portion between 28s and 40s for each one of the productions was manually selected with Praat to serve as the speech stimuli for the rating test. The selected speech fragments were decided to start from the moment the speaker starts telling the story and to not stop before a complete unit of speech is realised. Nonetheless, for some of the

audio files it was not possible to extract a sample that would fit the given duration criterion, since a complete speech unit would occur either too soon or too late. Thus, the files in reference were excluded from the evaluation task; however, three of them were chosen to be used as trials.

Additionally, because the data were not collected to serve the purpose of this study, a lot of interruptions were found to fill the data made by the researcher conducting the task. The quality of the sound was also compromised by a lot of background noise. All audios were consequently denoised and normalised for peak intensity through Praat and the interruptions by the researcher were manually removed. However, a considerable number of audios had to be discarded either because the interruptions occurred on the speakers' speech making it impossible to cut or because the sound remained quite heavy even after denoising and normalizing. In total 68 speech samples were finally presented to the judges in a unique randomized order using Praat.

Questionnaire for the judges

The questionnaire handed to the judges consisted of questions with regards to the judges' personal information and linguistic background. An emphasis was given on the judges' familiarity with Spanish/Catalan and Spanish-accented Greek. At the end of the questionnaire four questions on comprehensibility and fluency definition/perception as well as on the main features they paid attention to while making their judgments were included for qualitative analysis. The questions were directly taken from Suzuki and Kormos (2019) and were translated as close as possible to Greek. The questions as worded in Suzuki and Kormos's article (2019) are presented below:

• How would you define comprehensibility/fluency in your own words?

• What kinds of features did you pay attention to when you were rating comprehensibility/ fluency?

c. Procedures

L2-speech elicitation task

The data collection for the current MA thesis was part of a larger data collection (written and oral) for the purposes of the LETEGR2 project, whose objective is to investigate the way Greek is learned and taught in different learning contexts (second and foreign language context), as well as to identify good teaching practices that could potentially facilitate the teaching of Greek in each of these contexts. The data that is used here was collected in Barcelona, Spain at the beginning of a nine-month course (October) in 2019. It was carried out by one researcher and three trained researcher assistants. The data that are used for the present MA thesis were collected with each L2 speaker individually. A digital recorder was placed next to the participants as they were doing the picture narrative task. Participants were not provided with preparation time and they were not allowed to take any notes. At the end of the process a questionnaire was also administrated to elicit participants' personal information and language background. Before data collection, all participants were asked to sign a consent form for their participation in the study.

Rating procedure

The rating procedure took place in Athens, Greece from late December 2019 to early January 2020 within a three-week period. All judges evaluated comprehensibility in the first session and fluency in the second. The time gap between the two sessions was one day unlike the one-week gap in Suzuki and Kormos's study. A 9-point scale was used in both comprehensibility and fluency rating tasks (1=hard to understand/ not fluent at all, 9=easy to understand/very fluent correspondently). All judges were vaguely aware

of the story they were going to listen beforehand and in order to avoid familiarization effects they practiced on 3 speech samples, that served as trials and were not included in the analysis. During the trials judges were able to ask questions, that would clarify any doubt they might had, and were also allowed to set the sound volume as they wished. The sessions took place either in the judge's house or in the researcher's house. However, in the room there was no one else but the participant and the researcher and conditions of absolute silence were secured. All judges used the same pair of earphones and laptop to do the tasks. The instructions for both comprehensibility and fluency were very brief and given in Greek. Meanwhile, no definition was given, because we wanted to investigate native speaker's intuitive judgments on the two constructs. The instructions provided for comprehensibility followed the example of Derwing and Munro (2013), as presented in their appendix, while for fluency participants were encouraged to judge based on their own concept of fluency without providing them with any additional information. They were also encouraged to use the whole 9-point scale in both sessions. At the end of the first session judges were asked to complete the background questionnaire and to answer the two questions on comprehensibility. At the end of the second session they were asked to answer the correspondent two questions on fluency.

V. Analysis

1. Linguistic analysis

The speech samples were transcribed and linguistically analysed for a variety of features. The transcriptions were later pruned and analysed in terms of Analysis of Speech Units (AS-unit) and clauses according to Foster et al. (2000) guidelines for the third level of application of the unit. Additionally, to establish the reliability of the

analysis 25% of the data was randomly selected and analysed by another researcher. Cronbach's alpha was relatively high for both AS-units and clauses (α = .913 and α = .921; respectively). Moreover, following Segalowitz's distinction (Segalowitz, 2010) measures were obtained for speed fluency, breakdown fluency and repair fluency as well as for lexical complexity, structural complexity and accuracy covering most aspects of the three CAF components.

a. Fluency

Speed fluency

Initially two measures of speech rate (pruned, unpruned) were calculated for speed fluency. However, after further consideration it seemed reasonable to only include articulation rate for two reasons: i. articulation rate is a more accurate measurement of speed because all pauses are removed and it does not include breakdown fluency avoiding collinearity, ii. articulation rate was the measure that was used in the original article by Suzuki and Kormos (2019) and served comparability purposes better. Articulation rate was calculated by dividing the total number of syllables to the total phonation time, that is to say the sample's total duration minus the duration of the pauses (Kormos & Dénes, 2004; Suzuki & Kormos, 2019).

Breakdown fluency

Six measures were obtained to calculate breakdown fluency taking into account the frequency, the location and the duration of the pauses following Suzuki and Kormos's (2019) original article. As unfilled pauses were not calculated pauses shorter than 250m/s following the majority of the literature. The breakdown fluency measures are presented below:

Filled pause ratio: the total number of filled pauses was divided by the total number of words

Unfilled pause ration: the total number of silences was divided by the total number of words

Within-clause pause ratio: the total number of silences within clauses was divided by the total number of words

Between-clause pause ratio: the total number of silences between clauses was divided by the total number of words

Within-clause pause duration: the total duration of silences within AS-units was divided by the total duration of the sample

Between-clause pause duration: the total duration of silences between clauses was divided by the total duration of the sample

Repair fluency

One measure of repair fluency was calculated: dysfluency ratio per min (the total number of disfluencies was divided by the total duration and multiplied by sixty) (Skehan, 2003; Tavakoli & Skehan, 2005, Suzuki & Kormos, 2019).

b. Complexity

Lexical complexity was analysed using Guiraud index to calculate the type token ratio of each speaker in order to be able to control for text length effects (Gilabert, 2007). CLAN program was used in order to have a quick and precise account of the number of words and the times it appears in the text. Guiraud index was calculated in excel after manually adjusting the number of types and eliminating some items.

As for structural complexity, following calls in the literature (Bulté & Housen, 2012; Norris & Ortega, 2009;) to go beyond subordination, we looked at book subordination and overall structural complexity by looking at: i. Mean number of clauses per AS-unit (the total number of clauses was divided by the total number of AS-units) (Suzuki &

Kormos, 2019; Tavakoli & Foster, 2008; Tavakoli & Skehan, 2005), and ii. Mean length of AS-Units (the total number of words was divided by the total number of AS-Units) (Suzuki & Kormos, 2019; Tavakoli & Foster, 2008).

c. Accuracy

For the calculation of accuracy two categories of errors were created: i. Morphosyntactic errors, and ii. Lexical errors. As morphosyntactic errors were considered errors with regards to: inflection, declination, verb-subject agreement, gender, construction of subordinate clauses, aspects of mood and case selection. As lexical errors were calculated: wrong word choice, replacement of a word by its English or Spanish version, and word substitution. The mean number of the errors per 100 words was calculated for each one of the categories providing us with the morphosyntactic and lexical error rate. For the calculation of accuracy all self-repairs were not taken into account provided that the speaker's final choice was target-like. A 25% of randomly selected data was coded by another coder and Cronbach's alpha was calculated for both dimensions of accuracy (a = .889 for Morphosyntaxis, and .785 for Lexis). Considering that accuracy scores rarely reach absolute agreement among the coders, both values were accepted following Larson-Hall's suggested benchmark of .70-.80 (Larson-Hall, 2010).

2. Statistical analysis

In our first research question the relationship between perceived fluency and comprehensibility scores was addressed. The reliability across judges was sought for both constructs and values of Cronbach's alpha were found to be quite high (α =.878 for comprehensibility and α= .890 for perceived fluency). Thus, mean scores of comprehensibility and perceived fluency were computed for each speaker and were used as variables in the analysis. Descriptive statistics were next obtained and normality

of distribution was checked using the Kolmogorov-Smirnov test of normality. The assumption of normality was violated for the concept of comprehensibility but no outliers were found. Therefore, we resorted to non-parametric tests in order to proceed with the analysis. Spearman's Rank Order Correlation test was subsequently applied to calculate the strength of the relationship between the two concepts. Lastly, a Wilcoxon Signed Rank Test was run in order to check the statistical significance of the difference between the two sets of scores.

Furthermore, in order to investigate our second and third research questions, which were concerned with the linguistic correlations of comprehensibility and perceived fluency, a series of correlations and regression analysis models were employed. Descriptives statistics of the CAF measures were obtained and normality of distribution was checked using again the Kolmogorov-Smirnov normality test. Most of the variables did not follow the assumption of normality and displayed with outliers. It was decided to deal with the outliers by replacing them with a substitute value that resulted by calculating 2.5SD (standard deviation) above/below the mean. However, most of the variables were still presented with outliers. Therefore, we replaced the mean with the median and the standard deviation with the median absolute deviation (MAD) and we calculated 2.5MAD above/below median, following Leys et al.'s (2013) recommendation for robustly detecting the outlier boarders. Once the outliers were treated, descriptive statistics and normality tests were retaken. Normality of distribution was restored for some of the variables but yet not for all of them.

Moreover, Spearman's rho correlations were conducted between all CAF measurements and comprehensibility and perceived fluency scores. All CAF variables that were proved to have no correlation with each of the two constructs, were left out in the follow-up regression analysis. Five multiple regression models were created for

each dependent variable. The first three of them were used as preliminary analysis and employed a different set of CAF measurements to serve as the predictors based on the distinction among the three CAF dimensions, that is to say complexity, accuracy and fluency. It was therefore possible to significantly reduce the number of the independent variables in the final multiple regressions that were applied. There were chosen from the preliminary analysis only those variables that were shown to significantly explain (ps < .05) part of the variance in the scores of the dependent variables.

VI. Results

a. Research question 1

Descriptives on comprehensibility and perceived fluency showed that overall comprehensibility elicited higher scores than perceived fluency. Mean score for comprehensibility was 6.1, while for perceived fluency 5.12. Based on minimum and maximum values we could say that all speakers were moderate to highly comprehensible (min = 4.125, max = 8.875), while they were perceived somewhat less fluent (min = 3.125, max = 8.25).

The results of Spearman's correlation between the two subjective measures of comprehensibility and perceived fluency revealed a strong relationship among the two. Comprehensibility and perceived fluency significantly correlated ($r = .877$, $p < 0.01$) sharing more than 76% of their variance.

Moreover, the difference between the scores of comprehensibility and fluency, as depicted in the descriptives (Table1), shows that perceived fluency values (mean, minimum and maximum) are consistently lower than the ones of comprehensibility.

Table 1:Descriptives of comprehensibility and perceived fluency

	Mean	SD	Minimum	Maximum	Range
Comprehensibility	6.10294	1.178346	4.125	8.875	4.750
Fluency	5.12316	1.241669	3.125	8.250	5.125

The Wilcoxon Signed Rank test that was next applied confirmed that comprehensibility was evaluated by the judges significantly more leniently than fluency was (Mdn= 6 vs Mdn = 4.875, respectively; $z = -7.067$, $p < .001$).

a. Research questions 2 & 3

With regards to the Spearman's correlations between CAF variables and the two structures of comprehensibility and perceived fluency, the outcome of the analysis demonstrated that most of the linguistic variables correlated with both concepts (Table 2). As it turned out, the strongest correlation for both constructs was found to be within-clause pauses ratio ($r =-.744$, $p <.05$ and $r =-.741$, $p <.05$; for comprehensibility and perceived fluency respectively) sharing about 55% of their variance with each. The two accuracy measurements, Guiraud index, articulation rate and unfilled pause ratio exhibited as well significant moderate correlations higher than .5. However, words per AS-unit, between clause pause ratio, and dysfluency rate were not at all correlated with neither of the concepts. Interestingly enough, filled pause ratio and between-clause pause duration did not correlate with comprehensibility but were found to share a weak but significant negative correlation with perceived fluency ($r =-266.$, $p =.028$ and $r =-.261$, $p =.032$; respectively).

Finally, the outcome of the first three multiple regression models that were applied per CAF dimension led to a fourth regression model (Table 3), same for both concepts, that included the independent variables of: i. morphosyntactic error rate, ii. lexical error

rate, iii. Guiraud index, and iv. within-clause pause ratio. The model itself significantly explained 65,7% of the variance of comprehensibility and 66.9% of perceived fluency. There was no evidence of strong collinearity with VIF values ranging from 1.157 to 2.523. Nonetheless, it was observed that, even though beta value demonstrated within-clause pause ratio to be the strongest predictor of comprehensibility, the unique contribution of morphosyntactic error rate was higher ($r = -.298$, $p < .05$ and $r = -.245$, $p = .001$; respectively). A supplementary Spearman's inter-correlation across these four measurements found a moderate close to strong negative correlation between Guiraud index and within-clause pause ratio ($r = -.633$, $p < .05$), indicating that including lexical complexity in the model could have possibly reduced the unique contribution of breakdown fluency.

A fifth and final regression model (Table 4) was subsequently applied for comprehensibility and perceived fluency after excluding Guiraud index. The model itself loses a little bit of its total explanatory power explaining 63% of the variance of comprehensibility and 62.9% of perceived fluency. However, the unique contribution of the three remaining variables is far clearer. For the concept of comprehensibility within-clause pause ratio has a unique contribution of 17.64%, followed by morphosyntactic error rate (9.06%), while lexical accuracy was found to have a statistically insignificant unique contribution ($p = .243$). As for perceived fluency, all three predictors contributed significantly. Breakdown fluency had the strongest predictive power and a unique contribution of 15.68%. Grammatical accuracy was next (7.95%), while the unique contribution of vocabulary was a small but significant 2.85%.

To sum up, both comprehensibility and perceived fluency correlated with the same variables with the exception of filled pause ratio and between-clause pause duration that correlated with perceived fluency but not with comprehensibility.

Comprehensibility scores were found to be better predicted by within-clause pause ratio, morphosyntactic error rate and Guiraud, while perceived fluency was stronger associated with morphosyntactic error rate, within-clause pause ratio, Guiraud and lexical error rate. However, results changed a little bit once Guiraud was removed from the model.

Table 1:Spearman's rho Correlations of CAF measurements with Comprehensibility and Perceived Fluency

	Comprehensibility		Fluency	
CAF Measurements	r-value	p-value	r-value	p-value
Accuracy				
Morphosyntactic_Error_Rate	-.615**	.000	-.605**	.000
Lexical_Error_Rate	-.540**	.000	-.596**	.000
Lexical complexity				
Guiraud index	.599**	.000	.642**	.000
Structural complexity				
Nodes_per_ASunit	.293*	.015	.262*	.031
Words_per_ASunit	.077	.533	.046	.711
Speed fluency				
Articulation_Rate	.506**	00	.540**	.000
Breakdown fluency				
Filled_Pause_Ratio	-.213	.081	-.266*	.028
Unfilled_Pause_Ratio	-.534**	.000	-.529**	.000
Within_Clause_Pause_Ratio	-.744**	.000	-.741**	.000
Between_Clause_Pause_Ratio	.079	.521	.018	.881
Within_Clause_Pause_Duration	-.364**	.002	-.398**	.001
Between_Clasuse_Pause_Duration	-.139	.257	-.261*	.032
Repair fluency				
Dysfluency_Rate	-.188	.126	-.184	.133

It is worth mentioning that the total R squared value for the models does not equal all the squared part correlation values added up leaving out 35.49% in the case of

comprehensibility and 36,42% in the case of perceived fluency. This is probably because the part correlation values represent only the unique contribution of each variable, with any overlap or shared variance removed, while the total R squared value includes the unique variance explained by each variable and also that shared.

Table 3: Results of regression models for comprehensibility and perceived fluency

	R^2	Sig	Beta	Sig	Part	Tolerance	VIF
Comprehensibility model	.657	.000					
Morphosyntactic_Error_Rate			-.321	.000	-.298	.864	1.157
Lexical_Error_Rate			-.111	.239	-.088	.628	1.592
Guiraud index			.225	.030	.164	.529	1.889
Within_Clause_Pause_Ratio			-.389	.001	-.245	.396	2.523
Perceived Fluency model	.669	.000					
Morphosyntactic_Error_Rate			-.300	.000	-.279	.864	1.157
Lexical_Error_Rate			-.210	.025	-.166	.628	1.592
Guiraud index			.276	.007	.201	.529	1.889
Within_Clause_Pause_Ratio			-.291	.014	-.183	.396	2.523

Table 4: Regression models if Guiraud index is eliminated

	R^2	Sig	Beta	Sig	Part	Tolerance	VIF
Comprehensibility model	.630	.000					
Morphosyntactic_Error_Rate			-.324	.000	-.301	.864	1.157
Lexical_Error_Rate			-.113	.243	-.090	.628	1.592
Within_Clause_Pause_Ratio			-.541	.000	-.420	.601	1.663
Perceived Fluency model	.629	.000					
Morphosyntactic_Error_Rate			-.304	.000	-.282	.864	1.157
Lexical_Error_Rate			-.213	.030	-.169	.628	1.592
Within_Clause_Pause_Ratio			-.477	.000	-.370	.396	2.523

Finally, some indications coming from a short qualitative analysis of the questionnaires are reported hoping that they might assist to get a better insight. The judges' answers on the features they paid attention to while evaluating comprehensibility and perceived fluency were analysed on the basis of seven categories: pronunciation, grammatical accuracy, vocabulary appropriateness, speed, fluidity, pauses, repair fluency. The percentages of the individuals who referred to each of these categories is reported below (Table 5).

Table 2:Percentage of judges who directly referred to each of the categories

Category	**Comprehensibility**	**Perceived fluency**
Pronunciation	62.5	25
Grammatical accuracy	37.5	50
Vocabulary appropriateness	12.5	25
Speed	37.5	25
Fluidity	50	37.5
Pauses	37.5	25
Repair fluency	25	12.5

VII. Discussion

a. Research question 1

The first goal of this Master thesis was to examine the relationship between comprehensibility and perceived fluency and to draw reliable conclusions on whether these two concepts are distinguishable in the case of the 8 naïve Greek judges, who participated in this study. As expected, the results of the correlation test confirmed that a strong positive correlation does exist ($r = .877$, $p < .01$). Evidence of a strong relationship between these two concepts has been found before in the literature. With regards to the original article by Suzuki and Kormos (2019), the correlation that we found is smaller than the one they reported ($r = .95$, $p < .001$). This could be attributed to the different type of task (narrative descriptive vs argumentative) employed in the two studies. It has been reported before that narrative descriptive tasks tend to elicit lower scores of perceived fluency in comparison to other monologic or dialogic tasks (Derwing et al., 2004).

To a certain degree a relationship between comprehensibility and perceived fluency is inevitable, and therefore expected. A speech perceived as highly fluent is by default perceived as highly comprehensible as well (Suzuki & Kormos, 2019). Fluency presupposes comprehensibility in the same way that comprehensibility presupposes intelligibility. Nonetheless, there is evidence from the statistical analysis that advocates for the independency and the distinguishability of the two constructs. Descriptives showed that fluency scores were consistently lower than comprehensibility scores and the outcome of the non-parametric test confirmed the statistical significance of this difference (Mdn = 6 vs Mdn = 4.875, respectively; $z = -7.067$, $p < .001$). The two concepts were also distinguishable in Suzuki and Kormos (2019) according to the

results of the paired sample t-test they applied, despite that the exceptionally high correlation they found could attest for the opposite. The reported severity in perceived fluency evaluations by both studies might indicate that judges perceive their comprehensibility score as a purely descriptive indicator of the way they experience speech comprehension in terms of difficulty (Crowther et al., 2018), while they think of the fluency score as a qualitative indicator of performance evaluation. Subsequently, they tend to be more rigid when assessing fluency.

Moreover, perceived fluency shared a small but significant correlation with filled pause ratio and within-clause pause duration (r = -266., p =. 028 and r = -.261, p = .032; respectively), while comprehensibility did not. It could be the case that, if some of the judges viewed fluency as a synonymous to overall proficiency (Chambers, 1997; Tavakoli & Hunter, 2018), there were phenomena of oral performance that were judged negatively even though they did not impede comprehension importantly. If this is true, then this finding could possibly reinforce the idea that judges do differentiate between comprehensibility and perceived fluency. We can speculate in the qualitative analysis that features that are reported repeatedly by the judges when evaluating one construct lose their weight when they evaluate the other construct. A very straightforward example is that of pronunciation, although it has not been investigated in the article. In the case of comprehensibility assessment, pronunciation was actually reported by 62.5% of the judges rendering it the most frequently mentioned feature. This percentage falls radically (25%) in the case of perceived fluency. It is for certain impossible to have suddenly gone unnoticed by the judges when fluency was assessed. When a feature is mentioned, for whatever the reason, is by definition a feature that has been noticed. It is rather possible that judges consciously chose to ignore pronunciation inaccuracies when they were asked to evaluate fluency because they probably felt it to be less

significant or even irrelevant according to their personal definition and perception of the concept. It is indicative that the same inventory of features was reported for both comprehensibility and perceived fluency only with different frequencies. There was not one category that was mentioned for one of the constructs and not mentioned at least once for the other as well. Speech is a phenomenon perceived comprehensively but the qualities that can be detected by human ear are stable and unchangeable. What changes is the raters' perspective on the relative value of each feature. We would support that listeners are capable of attributing different weight on the temporal and linguistic features they notice according to which concept they evaluate each time.

b. Research question 2 & 3

The second and third research question of the paper sought to explain the way the three CAF dimensions are associated with comprehensibility and perceived fluency. Based on the correlation analysis a first observation is that both comprehensibility and perceived fluency correlated with most of the CAF measurements. Overall, the results of the correlations were very similar for comprehensibility and perceived fluency with minor differences. Accuracy, lexical complexity, nodes per AS-unit, articulation rate and measurements of breakdown fluency shared a stronger or weaker correlation with both concepts. With regards to comprehensibility, previous research has suggested that listeners depend on multiple linguistic dimensions and resources to make their judgements (e.g. Crowther, Trofimovich, Saito & Isaacs 2018; Saito, Trofimovich, & Isaac, 2017), and the same could be stated about perceived fluency as well.

It is noteworthy that listeners were sensitive to pause type and pause location when rating both constructs. The frequency of unfilled pauses was noticeably more important than that of filled ones ($r = -.213$, $p = .81$ and $r = -534$, $p < .01$ for comprehensibility; $r = -.266$, $p = .028$ and $r = -.526$, $p < .01$ for perceived fluency), while within-clause pause

ratio was the strongest correlate for comprehensibility as well as for perceived fluency sharing approximately 55% of its variance with each. The effect of breakdown fluency overall is supported by the qualitative analysis too. Pauses were directly mentioned by 37.5 % and 25% of the judges for comprehensibility and perceived fluency respectively. Moreover, if you bring this percentage together with the references to fluidity, that is most commonly linked to pausing behaviour, then the total percentage of judges who noticed breakdown fluency reaches up to 85% for comprehensibility and 62.5% for perceived fluency.

Furthermore, with regards to comprehensibility, within-clause pause frequency was also the strongest predictor as revealed by the multiple regression analysis. Pauses within clauses disrupt the sequencing of the message and it is reasonable to assume that it demands more cognitive effort by the listener to keep track of what was said and to eventually extract meaning from speech. As for perceived fluency, the results side with the line of research that attributes to pause behaviour an overall significant contribution to perceived fluency ratings (Bosker et al.,2013; Saito et al., 2018;). Mid-clause pause frequency was also the strongest predictor of perceived fluency in the original article by Suzuki and Kormos (2019), while our regression analysis revealed it to be one of the strongest predictors, second only to morphological error rate when Guiraud index was included in the model. There is evidence that the location of the pauses signals difficulties in different aspects of speech production. Pauses between clauses are related to problems in conceptual planning and are also common in native speakers, while mid-clause pauses indicate problems in linguistic encoding and lexical retrieval and are less commonly made by natives (Kormos, 2006; De Jong, 2016), and therefore are not perceived as natural.

Additionally, the role of grammatical accuracy, measured here as errors per 100 words, was verified as affecting comprehensibility and perceived fluency alike. It was found to be one of the two strongest predictors in the last two regression models (with and without Guiraud index) that were applied with a relatively stable unique contribution of about 9% for comprehensibility and 8% for perceived fluency. Judges' answers to the questionnaire at the end of the rating process also confirmed that grammatical accuracy was taken into consideration by 37.5 % when evaluating comprehensibility and by 50% of them when evaluating fluency. On the other hand, lexical error rate even though it moderately correlated with comprehensibility ($r = -.540$, $p < .001$), had no significant predictive power over it ($\beta = -111$, $p = 239$). This finding does not align with previous research that attributes significant value to both grammar and lexis in comprehensibility assessments (Crowther et al., 2018; Saito, Trofimovitch et al., 2016; Saito, Webb et al., 2016). Suzuki and Kormos (2019) as well suggest that lexical appropriateness is important in a picture descriptive speech elicitation task where a set of vocabulary items is predetermined. They further support that for an argumentative task like the one they used morphological features are far more crucial because of the lack of visual information. We would agree in that lack of a visual stimuli requires more attention to be paid in morphology rather than lexis especially when the content of the speech is not predetermined. In our study raters did not have any visual information either, but they got to know the story already by the end of the three trials, and to this extent the content of the speech was predefined for them in a way. However, Greek is a much more complex language in terms of morphology and syntax compared to English. The associations between the elements of an utterance are stronger and the simple reference to an element mentioned before requires the agreement of gender, number and case. It could therefore be the case that in morphologically more complex

languages grammatical accuracy influences comprehensibility more than lexis even in a picture description task. Further research is required in this direction to infer reliable conclusions.

What is more, the importance of lexical complexity has not been reported by Suzuki and Kormos (2019). In their original article only one of the three measurements of lexical complexity they took (lexical density) shared a medium correlation with comprehensibility ($r = -.543$, $p < .001$) and with perceived fluency ($r = -.551$, $p < .001$). However, the role of lexical complexity in our dataset has been decisive for the results of the regression analysis models. Guiraud index was detected to significantly alter the contribution of breakdown fluency. Its shared variance with within-clause pause ratio changed the balance of both comprehensibility and perceived fluency models. Once lexical complexity was removed the unique contribution of breakdown fluency was almost doubled for comprehensibility and perceived fluency alike. In the case of perceived fluency mid-clause pausing also changed to be the strongest predictor in place of morphosyntactic inaccuracy. Moreover, although values of VIF indicated no intercollinearity among the independent variables, the sum up of their unique variances never reached the total R^2 value of the model implying that there are overlaps between them. These results probably indicate that the relations between the different dimensions of CAF are interdependent in very complex ways. The presence of one phenomenon could create a domino effect to all the rest. The sole presence of breakdown fluency gives to the speaker fewer chances to develop lexical complexity. It could also influence accuracy by making the speaker lose track of what he was saying before affecting speech production in a similar way that effects speech processing, especially in cases where a level of automatization has not been acquired yet. It could also be the opposite. Problems in lexical retrieval or grammar will inevitably cause

pausing phenomena in speech. Speakers who lack linguistic knowledge will need more time to think and formulate their message. The relationship between linguistic and temporal aspects of speech is quite dependent. It requires great experience with language for a speaker to be able to strategically use the available resources to compensate for the lack of linguistic knowledge without affecting the temporal features of speech.

VIII. Conclusions

In the present study the relationship and distinguishability of comprehensibility and perceived fluency were investigated together with the unique associations that each one of them shares with measurements of CAF dimensions. We concluded that although these concepts correlate strongly with each other they still remain distinguishable. It was also suggested that judges assign to perceived fluency scores a truly evaluative value of oral performance and that they are probably capable of assigning different relative wight to each CAF variable based on which of the two constructs they are asked to rate. Furthermore, comprehensibility was best predicted by within-clause pause ratio, morphosyntactic error rate and lexical complexity but not by lexical accuracy contrary to previous literature findings. We assumed that the effect of morphosyntax on comprehensibility in the case of morphologically complex languages such as Greek is more significant than that of vocabulary even in tasks with predetermined lexical items. As for perceived fluency, morphosyntactic error rate, within-clause pause ratio, lexical complexity and lexical error rate were the strongest predictors. Overall, it was concluded that comprehensibility as well as perceived fluency scores depend on multiple variables of temporal and linguistic features. Finally, a reference to the

application of the CAF framework of analysis in the case of L2-Greek is considered necessary. We could say that overall following the guidelines that have been established for English, no major problems were encountered. However, the coding of accuracy was proved to be rather challenging as in many languages. In general, we would suggest that it is reasonable and practical to code morphology and syntax together. With regards to complexity and fluency, all CAF measurements were applied without any problems at least on the kind of data we had to deal with. It is important though, CAF measurements to be applied on native speakers' speech so that the specific patterns of L1-Greek speaking performance to be defined.

IX. Limitations & avenues for further research

The aim of this study was to partially replicate the original article by Suzuki & Kormos (2019) on the linguistic dimensions of comprehensibility and perceived fluency and to observe how results differ in a different population of participants. To a great extent the outcome of the present Master thesis replicated the results of the original article reinforcing the reliability and generalizability of some of its main findings. Moreover, comprehensibility and perceived fluency studies had never been conducted before with L2- Greek as far as we know. We hope to have made an important step and more studies to follow this line of research in the future.

Nonetheless, it is important to refer to the limitations of the present study. First of all, following Suzuki and Kormos (2019) we agree that comprehensibility and fluency perception should be investigated in different types of tasks and not only in picture/description that is the most common. Future research in L2-Greek speech perception should try to employ different types of tasks and also perhaps use different

levels of task complexity. Moreover, research should expand to different levels of proficiency from beginners to completely advanced students as well as to different language sets. In the case of L2-Greek the future research should include participants with different L1 in order to speculate the way crosslinguistic factors impact speech perception. In addition, it would contribute significantly to the field to work with bigger sample sizes even though the objective difficulties of doing so are well understood. Finally, it is important to secure in the future the optimal sound quality of the speech stimuli to avoid the influence of other confounding factors in the results. The special conditions of a Master thesis and the time constraints that it imposes were not favourable to this regard.

X. Bibliography

Abrahamsson, N., & Hyltenstam, K. (2008). The robustness of aptitude effects in near-native second language acquisition. *Studies in Second Language Acquisition, 30(4),* 481-509.

Andria, M. (2014). *Crosslinguistic influence in the acquisition of Greek as a foreign language by Spanish/Catalan L1 learners: The role of proficiency and stays abroad.* Unpublished doctoral dissertation. University of Barcelona

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,* 159–175.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA,* 23-46.

Chambers, F. (1997). What do we mean by fluency? *System*, *25(4),* 535-544.

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.

Crowther, D., Trofimovich, P., Saito, K., & Isaacs, T. (2018). Linguistic dimensions of L2 accentedness and comprehensibility vary across speaking tasks. *Studies in Second Language Acquisition, 40*(2), 443-457.

Crystal, D. (1987). The Cambridge Encyclopedia of Language. Cambridge University Press, Cambridge.

De Jong, N. H. (2016). Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching, 54*, 113–132.

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

DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition, 22,* 499–533.

Derwing, T.M., & Munro, M.J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching, 42,* 476-490.

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., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning, 54,* 665 – 679.

Flege, J., Munro, M.J., & MacKay, I.R.A. (1995). Factors affecting a degree of perceived foreign accent in a second language. *Journal of the Acoustical Society of America, 97,* 3125-3134.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied linguistics, 21*(3), 354-375.

Gilabert, R. (2007). The simultaneous manipulation of task complexity along planning time and [+/-Here-and-Now]: Effects on L2 oral production. *Investigating tasks in formal language learning, 20,* 44-68.

Gilabert, R., Barón, J., & Levkina, M. (2011). Manipulating task complexity across task types and modes. *Second language task complexity: Researching the cognition hypothesis of language learning and performance*, 105-140.

Housen, A., Kuiken, F., & Vedder, I. (2012). *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA.* Amsterdam, The Netherlands: John Benjamins Publishing Company.

Housen, A., & Simoens, H. (2016). Introduction: Cognitive perspectives on difficulty and complexity in L2 acquisition. *Studies in Second Language Acquisition*, *38*(2), 163-175.

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility. *Studies in Second Language Acquisition, 34(3),* 475–505.

Kang, O., Rubin, D., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *Modern Language Journal, 94,* 554-566.

Kormos, J. (2006). *Speech production and second language acquisition*. Mahwah, NJ: Lawrence Erlbaum Associates.

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System, 32,* 145–164.

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).

Language Teaching Review Panel. 2008. Replication studies in language learning and teaching: Questions and answers, *Language Teaching, 41*, 1-14.

Larson-Hall, J. (2010). A guide to doing statistics in second language research using SPSS. New York, NY: Routledge/Taylor and Francis Group.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning, 40,* 387-417.

Lennon, P. 2000. The lexical element in spoken second language fluency. In Hedi Riggenbach (ed.), *Perspectives on fluency*, 25–42. Ann Arbor: University of Michigan Press.

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly, 39*, 369-377.

Levis, J. M. (2006). Pronunciation and the assessment of spoken language. In R. Hughes (Ed.), *Spoken English, TESOL and applied linguistics: Challenges for theory and practice* (pp. 245 – 270). New York: Palgrave Macmillan.

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764-766.

Muñoz, C. (Ed.). (2006). *Age and the rate of foreign language learning* (Vol. 19). Multilingual Matters.

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.

Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences, 12,* 237–241.

Pallant, J. (2005). SPSS survival manual: A step by step guide to using SPSS for windows (version 12). *New South Wales, Australia: Allen & Unwin*.

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,* 151–169.

Préfontaine., Kormos, J., & Johnson, D.E. (2016). How do utterance measures predict raters' perceptions of fluency in French as a second language? *Language Testing*, *33*, 53 –73.

Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *International Review of Applied Linguistics in Language Teaching*, *43*(1), 1-32.

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*, 593–617.

Saito, K., Trofimovich, P., & Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics, 37(2),* 217-240.

Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (2016). Lexical correlates of comprehensibility versus accentedness in second language speech. *Bilingualism: Language and Cognition, 19(3),* 597-609.

Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using Listener Judgments to Investigate Linguistic Influences on L2 Comprehensibility and Accentedness: A Validation and Generalization Study, *Applied Linguistics, 38 (4),* 439–462.

Schmidt, R. (1992). Psychological mechanisms underlying second language fluency. *Studies in Second Language Acquisition, 14(4),* 357-385.

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. New York: Routledge.

Skehan, P. (2003). Task-based instruction. *Language Teaching, 36,* 1-14.

Suzuki, S. & Kormos, J. (2019). Linguistic Dimensions of Comprehensibility and Perceived Fluency: An Investigation of Complexity, Accuracy, and Fluency in Second Language Argumentative Speech. *Studies in Second Language Acquisition,* 1-25.

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,* 330–349.

Tavakoli, P., & Foster, P. (2008). Task design and second language performance: the effect of narrative type on learner output. *Language Learning, 58* (2), 439-473.

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.239-273). Amsterdam, The Netherlands: John Benjamins.

Appendix A: Questionnaire for the judges translated in English

Questionnaire

Statement of confidentiality:

Your name and other information gathered in this study will not be disclosed to any persons other than the investigator and his collaborators, and will only be used for statistical purposes without reference to individual participants' personal information.

Participant's name/code: _____
Date and time: _____

Personal Information

- Age: _____
- Profession: _____
- Place of birth: _____
- Mother's place of birth: _____
- Father's place of birth: _____
- Your current place of residence: _____
- Previous place(s) of residence (where you have lived for at least a few months; indicate when and for how long): _____

_____

Language background

- Native language (NL): _____
- Mother's NL: _____
- Father's NL: _____
- Do you speak any other language fluently? _____

_____

Please answer the following questions

1. Rate your knowledge in Spanish; (tick the appropriate box)

1	2	3	4	5	6	7	8	9

I do not speak at all **I speak fluently**

2. Rate your knowledge in Catalan; (tick the appropriate box)

1	2	3	4	5	6	7	8	9

I do not speak at all **I speak fluently**

3α. Rate your familiarity with Spanish-accented Greek (tick the appropriate box)

1	2	3	4	5	6	7	8	9

Not at all familiar **Very familiar**

3β. Rate your familiarity with Spanish language in general (tick the appropriate box)

1	2	3	4	5	6	7	8	9

Not at all familiar **Very familiar**

4. Rate your everyday exposure to Spanish-accented Greek (tick the appropriate box)

1	2	3	4	5	6	7	8	9

Never **all the time**

5. Rate your everyday exposure to foreign-accented Greek (tick the appropriate box)

1	2	3	4		5	6		7	8	9

Never **all the time**

6. Assess your overall level of confidence when rating comprehensibility (tick the appropriate box)

1	2	3	4	5	6	7	8	9

Not at all confident **very confident**

7. Assess your overall level of confidence when rating fluency (tick the appropriate box)

1	2	3	4	5	6	7	8	9

Not at all confident **very confident**

8. Have you ever taught Greek as a second/foreign language (professionally or voluntarily)?

9. How would you define comprehensibility in your own words?

10. What kinds of features did you pay attention to when you were rating comprehensibility?

11. How would you define fluency in your own words?

12. What kinds of features did you pay attention to when you were rating fluency?

Appendix B: Instructions given to the judges before the rating tasks (translated in English)

For comprehensibility: You will hear second language learners of Greek describing a story about some kids and their dog who are getting ready to go to a picnic. All of the speech samples are taken from the first 30–40 seconds. What we would like you to do is make a judgment about each sample. We will ask you to say how easy or difficult the sample is to understand, using a 9-point scale. Please, try and use the whole scale while judging. You might be able to understand everything but it may require a lot of effort on your part—so what we are interested in is the effort you put in. Can you understand it without even thinking about it, or do you have to work at it? First, you will practice with three trials during which you can ask any question you want and set the volume as you prefer. Once the main task starts you cannot stop or ask any questions. (adaptation from Derwing & Munro, 2013).

For perceived fluency: You will hear the same speech samples you heard the other day. What we would like you to do this time is evaluate fluency using a 9-point scale. Please, try and use the whole scale while judging. Make your assessment based on your perception of the concept, whatever that might be. First, you will practice with three trials during which you can ask any question you want and set the volume as you prefer. Once the main task starts you cannot stop or ask any questions.

Appendix C: The preliminary regression models per CAF dimension

Table 3: Regression model 1: Accuracy

	R^2	Sig	Beta	Sig	Part	Tolerance	VIF
Comprehensibility model	.454	.000					
Morphosyntactic_Error_Rate			-.429	.000	-.410	.914	1.095
Lexical_Error_Rate			-.409	.000	-.391	.914	1.095
Perceived Fluency model	.492	.000					
Morphosyntactic_Error_Rate			-.396	.000	-.379	.914	1.095
Lexical_Error_Rate			-.474	.000	-.453	.914	1.095

Table 4: Regression model 2: Complexity

	R^2	Sig	Beta	Sig	Part	Tolerance	VIF
Comprehensibility model	.392	.000					
Guiraud index			-.598	.000	-.575	.925	1.081
Nodes_per_ASunit			-.084	.407	-.081	.925	1.081
Perceived Fluency model	.411	.000					
Guiraud index			-.645	.000	-.620	.925	1.081
Nodes_per_ASunit			-.014	.889	-.013	.925	1.081

Table 5:Regression model 3: Fluency

	R^2	Sig	Beta	Sig	Part	Tolerance	VIF
Comprehensibility model	.548	.000					
Articulation_Rate			.129	.343	.081	.392	2.554
Unfilled_Pause_Ratio			-.196	.156	-.122	.386	2.589
Within_Clause_Pause_Ratio			-.475	.007	-.234	.243	4.116
Within_Clause_Pause_Duration			-.088	.393	-.073	.679	1.472
Perceived Fluency model	.547	.000					
Articulation_Rate			.206	.133	.129	.392	2.554
Unfilled_Pause_Ratio			-.209	.131	-.130	.386	2.589
Within_Clause_Pause_Ratio			-.403	.022	-.198	.243	4.116
Within_Clause_Pause_Duration			-.102	.327	-.084	.679	1.472