Winter 1-31-2024

Segmental Error Gravity in L2 Arabic Speech Comprehensibility and Accentedness: L1 Consonant Age of Acquisition as a Predictor

Mark Papai
markpapai@aucegypt.edu

The American University in Cairo

School of Humanities and Social Sciences

Segmental Error Gravity in L2 Arabic Speech Comprehensibility and Accentedness: L1 Consonant Age of Acquisition as a Predictor

A thesis submitted to

The Department of Applied Linguistics

In Partial Fulfillment of the Requirements for

The Degree of Master of Arts in

Teaching Arabic as a Foreign Language

By Mark Papai

Under the supervision of

Dr. Raghda El Essawi

August 2023

Acknowledgements

First and foremost, I would like to extend my gratitude beyond the thesis work itself to the entire TAFL program for making this experience possible and enabling me to pursue my long-time dream of becoming an Arabic teacher. Special thanks to Dr. Zeinab Taha for always being flexible and accommodating with regards to the pathways and timeline of completing degree requirements. I also thank her for providing me with a solid theoretical foundation in linguistics during the first year of my study. I would like to express my gratitude to Dr. Raghda El Essawi for providing rigorous practical and pedagogical training during my second year of study. Her relentless pursuit of clarity, relevance and coherence in justifying pedagogical decisions has not gone fruitless. Lastly, I would like to thank Dr. Atta Gebril for inspiring me to draw comparisons between teaching English and Arabic through his class discussions. For Arabic teaching to succeed, such crosslinguistic comparisons are indispensable.

While I entered this program with a passion for teaching, I am now leaving with a passion for both teaching and research. Through getting introduced to the fields of applied linguistics and second language acquisition, I found my intellectual home. The concepts in the field have compelled me to ponder research directions that would enrich the Arabic teaching profession. This thesis represents the first step in this long agenda, allowing me to combine my passion for phonology with my background in psychology. I thank my supervisor, Dr. Raghda El Essawi and my first reader, Dr. Nihal Sarhan for always reminding me to keep in mind the pedagogical relevance of this investigation and not to overextend myself. Their comments have been crucial to strengthening my line of thought and any remaining errors are thus mine only.

To my partner: thank you for bearing with me during the research process and being there for me through the ups and downs. Your presence has meant the world to me.

Abstract

Second language (L2) pronunciation studies have found that the intelligibility (i.e., listeners' actual understanding) of L2 speech is most closely related to its comprehensibility (i.e., listeners' ease of understanding) rather than to its overall nativelike pronunciation. The segmental errors that are most detrimental to communication are predicted by phoneme Functional Load (FL): mispronouncing high FL segments affects speech comprehensibility more negatively than mispronouncing low FL ones. However, no data are available on the FL hierarchy of Arabic segments. On the other hand, FL correlates highly with consonant age of acquisition (CAoA) in languages that rely heavily on consonants to contrast meaning, and whose listeners rely on consonants to process speech. In these languages, the higher the FL of consonants, the earlier they are acquired by children. Arabic phonology and psycholinguistic data suggest a strong consonant bias, possibly meaning that CAoA could be used in place of FL to predict consonantal error gravity, with early-acquisition consonants representing high FL and late-acquisition ones representing low FL. To test this hypothesis, a speech perception study was conducted. Twenty-one native speakers of Egyptian Arabic listened to 23 words read aloud by ten L2 learners of Arabic containing either an early-acquisition consonant error, a late-acquisition consonant error, or no error, and rated them on comprehensibility and foreign-accentedness. Results suggest that early-acquisition consonant errors are more detrimental to comprehensibility and are perceived as more foreign-accented. In addition, a moderate-to-strong correlation was revealed between comprehensibility and foreign-accentedness. The findings provide the first empirical evidence that could be used to set instructional priorities in Arabic L2 pronunciation pedagogy.

Keywords: TAFL, consonant errors, mispronunciations, error gravity hierarchy, intelligibility, comprehensibility

Table of Contents

List of Tables and Figures

List of Abbreviations

AFL: Arabic as a foreign language

CAoA: consonant age of acquisition

ECA: Egyptian Colloquial Arabic

FL: Functional Load

L1: native language

L2: second language

MSA: Modern Standard Arabic

SLA: Second Language Acquisition

Chapter One: Introduction

In second language (L2) pronunciation, difficult does not necessarily mean important. Acquiring L2 phonology tends to be difficult for most learners, due to the interference of L1 pronunciation features, among other factors (Archibald, 2021). This difficulty means that most adult learners do not end up developing nativelike accents in a foreign language (Moyer, 2013). However, the presence of a foreign accent does not automatically mean that the speech is less understandable. Rather, listeners' actual understanding of foreign-accented utterances is more closely related to the amount of effort they have to put into understanding the utterance, a concept which has been termed *comprehensibility* in L2 pronunciation literature (Munro & Derwing, 2020). Certain aspects of a foreign accent cause more difficulty for the listener to decode the message, and therefore pronunciation instruction should prioritize such errors as more important to address (Levis, 2020). Pronunciation can be broken down into segmental (i.e., phonemes: consonants and vowels) and suprasegmental (e.g., stress, rhythm, intonation, tone) features. While there has been a debate about the relative importance of the correct production of segmentals vs. suprasegmentals in speech comprehensibility (Wang, 2022), attempts to determine a hierarchy of importance among different segments have been more successful (Sewell, 2021), through utilizing the Functional Load principle.

The chapter starts with an outline of the historical developments in the field of Second Language Acquisition (SLA) that led to the empirical findings on the error gravity hierarchy of segmental features. It then presents the previous attempt in Arabic as a foreign language (AFL) to determine such a hierarchy, discusses the underlying assumptions behind it and provides a more detailed discussion of that hypothesis. Finally, the chapter presents an empirical speech perception study to test this hypothesis in predicting L2 Arabic segmental error gravity.

Empirical evidence on pronunciation error gravity in Arabic is important for setting instructional priorities in AFL, since there is not enough time to address all mispronunciations in the classroom due to the difficulty involved in L2 phonological production. At the same time, pronunciation also cannot be ignored. Mispronunciations that decrease speech intelligibility and comprehensibility the most cause breakdowns in communication, avoidance on the part of interlocutors and thus hurt learners' chances of receiving input and engaging in interaction.

1.1. Background

The fields of Second Language Acquisition (SLA) and L2 pedagogy have gone through different waves of approaches to L2 pronunciation (Murphy & Baker, 2015). For a long time, L2 teaching and research were motivated (either explicitly or implicitly) by the *nativeness principle*: the premise that a nativelike accent is both achievable and desirable for L2 learners (Levis, 2005, 2020). While early approaches to foreign language teaching put overwhelming emphasis on getting rid of foreign accents in learners' speech, subsequent methods that emerged in the 70s and 80s largely neglected the explicit instruction of pronunciation, a neglect that still has lasting effects in classrooms (Derwing, 2017). The effect of a native-nonnative binary was evident in both research and teaching, with the pendulum swinging from the preoccupation with achieving nativelike, "accent-free" speech production, to the complete disregard of pronunciation in L2 classrooms due to emerging evidence for the futility of pursuing nativelikeness (Derwing & Munro, 2022). This limiting dichotomy was disrupted by the empirical findings of Munro and Derwing (1995a), whose seminal study established that foreign accents have little to do with the intelligibility (listeners' actual understanding) of L2 speech. Rather, it is comprehensibility (listeners' ease of understanding) that shows a closer relationship with speech intelligibility. Subsequent research then emerged to discover the factors affecting L2 speech comprehensibility

and intelligibility (Crowther et al., 2022), bringing about the acceptance of intelligibility as a guiding principle for L2 pronunciation pedagogy and a viable alternative to the inadequate nativeness standard (Levis, 2020).

Before the findings of Munro and Derwing (1995a), pockets of research emerged that were starting to investigate other aspects of L2 speech, related to how understandable that speech is. This was captured by two constructs: *comprehensibility* and *intelligibility.* The concepts have been objects of confusion in SLA literature (Trofimovich et al., 2022). As it is currently understood, comprehensibility is the ease with which a listener can understand the speaker's utterance, rated on (usually 9-point) Likert scales. On the other hand, intelligibility refers to the actual understanding of the utterance demonstrated by the listener, usually through transcriptions of the speech recording. Yet, early studies sometimes used the term comprehensibility for what we now refer to as intelligibility (measured as transcription accuracy), while others used intelligibility to refer to the listener's ease of understanding. Still others operationalized intelligibility in a way that was impossible to differentiate from accentedness, suggesting the pervasiveness of the notion that foreign accents are inherently more difficult to understand. Nevertheless, early studies conducted on the extent of understandability of an L2 utterance revealed important insights, showing, for example, that Foreigner Talk can be triggered by the reduced comprehensibility of L2 speech (Gass & Varonis, 1985; Varonis & Gass, 1982).

The need to dispel the notion that foreign-accentedness necessarily causes reduced speech intelligibility led to the first empirical investigation of the interrelationships between the constructs of accentedness, intelligibility and comprehensibility. In their seminal study, Munro and Derwing (1995a) showed that listeners' ratings of speakers' degree of accentedness (on 9-point Likert scales) displayed only a low correlation with their speech intelligibility, which was

measured through the accuracy of listener transcriptions. Instead, comprehensibility, also rated on a 9-point Likert scale measuring listeners' difficulty of understanding, was the better predictor of intelligibility scores. These results have since been replicated on languages other than English, including Spanish (Nagle & Huensch, 2020), and Arabic (Ali, 2023). At the same time, a different strand of research emerged measuring only accentedness and comprehensibility, leaving out intelligibility, possibly due to the methodological and statistical convenience of only including Likert scales in the study, and the fact that comprehensibility predicts intelligibility fairly reliably.

The findings that established the independence of intelligibility from overall accentedness led to the promotion of the *intelligibility principle* in L2 pronunciation, which states that the goal of L2 speaking instruction is for learners to produce speech that is understood with relative ease (i.e., both intelligible and comprehensible speech) by listeners (Levis, 2005, 2020). While originally proposed as an alternative to nativeness standards in lingua franca languages such as English, it is now an accepted pedagogical framework for addressing pronunciation in foreign language classrooms generally. The principle holds that those aspects of a foreign accent that decrease speech intelligibility and comprehensibility need more instructional focus. Consequently, the variables that contribute positively or negatively to L2 speech intelligibility (in its broad sense that encompasses comprehensibility) are important to research in any language. At the same time, L2 speakers are not considered responsible for minimizing any potential negative reactions that might occur in response to their foreign accents, and the onus is on listeners to adjust their attitudes and prepare themselves for foreign-accented speech (Derwing et al., 2002; Derwing & Munro, 2022). In this new framework, intelligibility is co-created by speaker and listener, who each do their part in ensuring communicative success.

Accordingly, numerous studies have since explored the factors that affect L2 speech comprehensibility and accentedness ratings (Crowther et al., 2022). The evidence most relevant to L2 pronunciation pedagogy comes from studies investigating the relative contribution of segmental and suprasegmental features (Wang, 2022), and the ones aiming to determine the relative weight of different segmental errors. The relative importance of segmental errors has been approached from the perspective of the communicative role that the segment plays in the target language, conceptualized as the Functional Load (FL) of a particular phoneme or phonemic opposition (Brown, 1988; Catford, 1987). In its most basic formulation, FL is the amount of contrastive work that a phonemic opposition performs to keep the meaning of words separate in a language. The FL hypothesis in SLA states that the higher the FL of a phoneme, the more detrimental its substitution is to speech comprehensibility (e.g., pronouncing "pin" as "bin" has the potential to disrupt communication more than pronouncing "think" as "sink", because the /p/-/b/ contrast has a higher FL in English than /θ/-/s/; meaning that the former pair has more minimal pairs than the latter). The early impressionistic calculations of English FL hierarchies have been borne out in subsequent empirical studies (Alnafisah et al., 2022; Munro & Derwing, 2006; Suzukida & Saito, 2021; Thir, 2020), and a recent investigation on Chinese as a second language has also confirmed the FL hypothesis (Bao et al., 2022). Overall, these results suggest that FL hierarchies can be robust in setting instructional priorities in L2 pronunciation.

1.2. Statement of Problem and Significance

Munro and Derwing (2015) call for the investigation of pronunciation error gravity hierarchies in languages other than English. The mispronunciations that hurt speech intelligibility and comprehensibility the most are important to uncover in any language, because they have the greatest potential to cause misunderstandings and breakdowns in communication. For example, it

has been shown that the reduced comprehensibility of non-native speech causes native interlocutors to simplify their own responses and also causes a reduction in the quantity of their speech. Interlocutors also appear to cut interactions short with non-native speakers whose speech is difficult to understand. This can seriously hurt L2 learners' chances for interaction and communicative practice, potentially preventing them from reaching high levels of ultimate attainment. For these reasons, such mispronunciations should receive selective instructional focus, which could be achieved through focus-on-form exercises, which integrate pronunciation in communicative activities the same way as vocabulary and grammar are integrated (Isaacs, 2009).

When it comes to Arabic, Hellmuth (2014) points out that there is no available information on the relative FL of segments. To circumvent this gap, she relies on evidence from L1 acquisition studies showing a correlation between FL and consonant age of acquisition (CAoA) to present a segmental error gravity hierarchy in Arabic L2 pronunciation pedagogy. Her recommendations are based on Amayreh and Dyson's (1998) speculations that some consonants acquired early by Jordanian children might have high FL, and on early findings by Stokes and Surendran (2005) showing that FL is strongly correlated with CAoA in English and Dutch. In these languages, the higher the FL of a consonant, the earlier it is acquired by NS children. However, before pedagogical recommendations can be made, they require empirical confirmation. First, the hypothesized relationship between FL and CAoA in Arabic needs to be examined in more detail. Then, CAoA data needs to be explored in multiple dialects, since acquisition of consonants in different dialects might not follow the same path.

The higher the correlation between FL and CAoA in a language, the more its speakers rely on consonants for speech processing, and the earlier they develop this consonantal bias as

children. Furthermore, the higher this correlation is, the more that language relies on consonants to contrast lexical meaning. For example, Spanish shows a very strong correlation between FL and CAoA, and accordingly, Spanish speakers show a consonant bias in speech perception that develops within the 1st year of life. Spanish also has 19 consonants, five vowels and no contrastive suprasegmental features. English shows a relatively weaker correlation between FL and CAoA, and English-speaking children only develop a consonantal bias in their 3rd year of age. The language has 24 consonants, 19–20 vowels, and contrastive stress. In contrast, Mandarin and Cantonese only exhibit a weak relationship between FL and CAoA. Likewise, Mandarin- and Cantonese-speaking children retain a vowel bias in their 3rd year of age, and do not end up developing a consonant bias in speech perception even into adulthood. These languages contain more vowels than consonants, and also 4–6 contrastive tones.

Following the logic above, there are reasons to believe that CAoA is strongly associated with FL in Arabic and therefore could be used to predict segmental error gravity. Firstly, the overwhelming majority of Arabic segments are consonants: 25–28 depending on the variety of the language, compared to only 6–8 vowels (three short and three to five long vowels, depending on the variety). The only contrastive suprasegmental features are gemination in word-medial and final position and vowel length. Secondly, the language uses consonants in tri- and quadriliteral roots that are inserted into vocalic patterns to form words. Whereas the vocalic pattern is used to express morphological class, the consonantal root carries semantic information. This has been shown to affect speech processing in both Arabic and Hebrew (which uses consonantal roots and vocalic patterns in the same way), making listeners favor consonantal information to identify words both individually and within a sentence (Aldholmi & Pycha, 2023; Lador-Weizman & Deutsch, 2022). In addition, a computational model has been found to be better able to identify

word boundaries in Arabic when fed consonant-only data compared to data containing both consonants and vowels (Kastner & Adriaans, 2018). The same model was less accurate in identifying English word boundaries when given consonant-only representations, indicating that children acquiring these two languages probably utilize consonants and vowels differently in speech perception and lexical learning, with a clearer consonant benefit in Arabic.

As for the acquisition order of Arabic consonants, we can only rely on large cross-sectional investigations that compare multiple age groups in the process of acquiring their native phonologies. These have been conducted on Jordanian (Amayreh & Dyson, 1998), Syrian (Owaida, 2015) and Egyptian Arabic (Elrefaie et al., 2021). While Jordanian shows different ages of acquisition for consonants, recent data on Syrian and Egyptian show similarities. Based on these, the most commonly mispronounced consonants in Arabic L2 speech can be grouped into early-acquisition and late-acquisition consonants, mirroring studies that compared the effects of high FL and low FL mispronunciations. The early-acquisition ones are the following: ع /ʕ/, ح /ħ/, ه /h/, and ء /ʔ/. The following consonants can be categorized as late-acquisition: خ /x/, غ /ɣ/, ض /dˤ/, ط /tˤ/, ظ /ðˤ/, ص /sˤ/, ق /q/.

1.3. The Present Study and Research Questions

Based on the evidence presented above, an empirical speech perception study was conducted to investigate the predictive power of CAoA with regard to consonantal error gravity in L2 Arabic speech. Forty words were read aloud by ten L2 speakers of Arabic and two NSs of Egyptians Arabic. Thirty of them belonged to the L2 speakers, with ten of these containing an early-acquisition consonant error, ten containing a late-acquisition consonant error, and ten containing no errors. Twenty-one NSs of Egyptian Arabic without linguistic training or teaching experience listened to the words on Qualtrics and rated each word on 9-point Likert scales

measuring comprehensibility and accentedness. Effects of CAoA on the two constructs were examined. Accordingly, the study aimed to answer the following research questions:

1. Does consonant age of acquisition predict the relative effect of consonantal errors on ratings of speech comprehensibility in L2 Arabic?
2. Does consonant age of acquisition predict the relative effect of consonantal errors on ratings of speech accentedness in L2 Arabic?
3. What is the relationship between L2 Arabic speech comprehensibility and accentedness for read-aloud words?

1.4. Definitions and Operational Variables

Comprehensibility: the degree of listening effort required to understand the speaker's utterance. In psycholinguistic terms, it captures processing fluency. In the current study, it is operationalized as a 9-point Likert scale, with "1" denoting "extremely easy to understand", and "9" denoting "impossible to understand", averaged for each error condition.

Accentedness: the degree of difference from a local pronunciation norm as perceived by the listener. In psycholinguistic terms, it represents the perceptual salience of the phonological features of the spoken utterance. In the current study, it is operationalized as a 9-point Likert scale, with "1" denoting "no foreign accent at all", and "9" denoting "heavy foreign accent", averaged for each error condition.

Segmental error: the substitution of a target language phoneme with another phoneme of the target language.

Functional Load: the amount of contrastive work performed by a pair of phonemes in a linguistic system. In terms of information theory, it is the amount of information loss that the linguistic

system would sustain if the phonological contrast were neutralized. It is not operationalized in the current study, as it is inferred through the proxy variable *consonant age of acquisition*.

Consonant age of acquisition: the age at which a consonant is correctly produced in all positions by 90% of children without developmental disorders. In the current study, it is operationalized as a categorical variable with two levels: early acquisition (before age 4) and late acquisition (after age 4).

1.5. Delimitations

There are several important factors with the potential to affect L2 speech comprehensibility, and accentedness that are nevertheless outside the scope of the current study, among them the effect of vowel and suprasegmental errors, lexicogrammatical errors, and the moderating effects of listener and speaker-based variables (for an annotated research timeline reviewing these factors, see Crowther et al., 2022). While vowels generally have lower FL than consonants, they are nevertheless acquired early by NS children (Nazzi & Cutler, 2019), meaning that FL is probably not a good predictor of vowel emergence. Thus, the current framework that is employed by this study is not adequate to address the effect of vowel errors. In addition, it might be difficult to directly compare consonant and vowel errors, due to the different position they assume within syllables. Secondly, lexicogrammar errors are best investigated in extemporaneous and spontaneous speech samples, which the controlled nature of the current study does not allow for. Listener and speaker factors can also be potentially investigated in future studies, after the effects of purely phonological errors have been triangulated in more controlled investigations. In general, the current study falls towards the more psycholinguistic end of SLA studies, due to the extent of experimental control involved.

Chapter Two: Review of Literature

This chapter includes a review of the literature on which the research questions are based. First, it discusses the sources of foreign accents and the problematic nature of pursuing nativeness standards in L2 pronunciation pedagogy. This is then followed by a review of relevant literature that establishes the partial independence of the constructs of intelligibility, comprehensibility and accentedness, and shows that comprehensibility is the better predictor of intelligibility compared to foreign-accentedness. Then, the segmental factors influencing speech comprehensibility and accentedness ratings are discussed. While there is a significant body of literature investigating the multitude of factors affecting speech comprehensibility and accentedness, the review is restricted to the empirical studies that a) investigate the contribution of segments (consonants and vowels); b) establish the predictive power of Functional Load in segmental importance; c) uncover the correlation between Functional Load and Consonant Age of Acquisition in Arabic; and d) provide data on Consonant Age of Acquisition in Arabic.

2.1. L2 Speech Comprehensibility and Accentedness

2.1.1. Accentedness

Accent is one of the most salient aspects of human beings, besides physical appearance. Listeners are able to recognize a dissimilar or unfamiliar accent within a matter of milliseconds (Flege, 1984). What is more, listeners can sometimes detect foreign accents in languages they are not proficient in (Major, 2007). Despite the common colloquial use of the word, everyone has an accent: the term refers to the systematic patterns of sound that someone's speech exhibits. However, accents have not been treated equally in the social realm: some have been historically afforded higher status and this has been reflected in foreign language teaching, too (e.g., through the privileging of inner circle English varieties for teaching English). While foreign accents can

sometimes elicit negative reactions from listeners, it is now widely accepted that accent reduction and elimination are not the right approach in L2 pedagogy, which favors the intelligibility principle in pronunciation instruction (Levis, 2005, 2020). The following paragraphs lay out the rationale for this shift in priorities and principles from the standpoint of foreign accent.

Historical approaches in linguistics and L2 pedagogy tended to treat foreign accents as something to eventually get rid of (Murphy & Baker, 2015). These approaches were imbued with *native speakerism* that holds the purported native speaker as the norm and the goal to strive for in language learning at every level of linguistic structure (Holliday, 2006). This ideology was exemplified to an extreme level by the Audiolingual Method, which set out to eradicate deviations from native phonological systems through incessant and repetitive drilling of speech patterns (Baker, 2017). While calls occasionally sprung up to invoke intelligibility as a principle for pronunciation, these did not gain enough currency to become dominant (Murphy & Baker, 2015). With the advent of communicative language teaching, the explicit treatment of pronunciation took a backseat, as evidence for the futility of pursuing nativelike accents emerged (Levis & Sonsaat, 2017). It was also thought that pronunciation was merely a function of proficiency and did not need to be addressed in a directed manner. This neglect is evidenced in the dearth of empirical research during what Murphy and Baker (2015) termed the third wave of pronunciation pedagogy. This shift from a nativeness standard to no explicit treatment of pronunciation is described by Levis and Sonsaat (2017) as a move from accuracy to fluency in instructional priorities.

The existence of foreign accents could be considered a manifestation of the exceptional difficulty learners face in L2 phonological production. The sources and causes of difficulty have

been conceptualized differently by various theoretical approaches (Archibald, 2021). Contrastive Analysis predicted positive transfer of L1 phonological features that are similar to the L2 ones and negative transfer of dissimilar features, making dissimilar phonological features of L2 more difficult to acquire (Archibald, 2017). Another approach has posited markedness as a source of difficulty in L2 phonological production (Eckman, 2008), in that more marked features are harder to acquire (e.g., consonant clusters). What is largely shared between these approaches is the assumption of fundamental difference in adult L2 phonological acquisition, which posits the existence of a critical period for native-like acquisition of accent. In addition, this fundamental difference is amplified in the case of phonology, compared to the acquisition of lexicon and morphosyntax, making nativelike attainment of L2 phonology an unrealistic standard (Caldwell-Harris & MacWhinney, 2023).

Not only is sounding nativelike an unrealistic goal, but it is also an unfair and unjust standard. Accents are powerful social markers of identity, and irrespective of difficulty, speakers might not wish to sound nativelike but want to retain their non-native accents in an attempt to signal their belonging to a certain group and to assert their identities. Adults are thought to have an established sense of identity that is tied to their native languages. Moyer (2013) lists some examples of reasons for L2 speakers wanting to project a non-native identity: maintaining an interesting personality, wanting to showcase that acquiring the L2 took hard work, and wanting to fit in with other L2 learner peers. These examples show that at times, regardless of ability, conscious choice plays a role in the manifestation of foreign accent and that this choice can be driven by social factors and personal preferences that need to be respected.

Foreign accents are but one example of non-standard language use and as such, are subject to dominant language ideologies (i.e., native speakerism) that stigmatize non-standard

varieties, whether native or non-native (Gluszek & Dovidio, 2010; Moyer, 2013). This has a host of ramifications including linguistic prejudice and even discrimination based on speech characteristics (Baugh, 2017). For example, callers judged as sounding black could be denied housing opportunities in the United States (Purnell et al., 1999). In a similar vein, Americans have been found to judge foreign accents lower on dimensions of status and solidarity compared to native ones (Dragojevic & Goatley-Soan, 2022), with an additional hierarchy between the non-native accents. Similar hierarchical attitudes to perceived non-standard speech characteristics have been observed in the case of Arabic (Gwasmeh, 2021). Nevertheless, as Munro and Derwing (2020) point out, L2 learners should not bear the burden of mitigating negative listener attitudes: the onus is on listeners to adjust to foreign-accented speech and training seems to be effective in this regard (Derwing et al., 2002).

Despite the overwhelming body of evidence and arguments against imposing native-like accents on L2 learners, it is still beneficial to measure the degree of perceived foreign accentedness in speech perception studies involving L2 speakers. Such measures can potentially tap into language attitudes and give us information about the features of speech that native listeners associate with sounding foreign. This information could be used to train native speakers to listen to non-native speech, similarly to what Derwing et al. (2002) carried out. Still, there is little use in measuring foreign accentedness in isolation, and it is best measured in connection with comprehensibility (and intelligibility), which will be discussed in the following section.

2.1.2. Comprehensibility

Comprehensibility can be thought of as the lowest common denominator in intelligible speech: highly comprehensible speech is likely to also be highly intelligible, making comprehensibility a practical and convenient measure in research even in the absence of direct

measurements of intelligibility. However, in certain cases, intelligible speech can still receive low comprehensibility ratings, which has ramifications on the success of the interaction. For this reason, the intelligibility principle in L2 pronunciation teaching states that students should target comfortably intelligible pronunciation, which entails high comprehensibility and high intelligibility (Levis, 2005, 2020). The overwhelming majority of studies that have investigated the correlation between foreign-accentedness and comprehensibility in numerous languages found only a moderate level of correlation, and this partial separation between the two constructs has been confirmed even at the level of individual words (Uchihara, 2022). The superiority of comprehensibility over foreign-accentedness in predicting intelligibility has been replicated in L2 Arabic, as well (Ali, 2023). The following sections discuss the historical, theoretical and methodological underpinnings of comprehensibility as a research construct in L2 pronunciation.

The study of speech intelligibility originates from the field of telecommunications, where researchers were originally interested in sound clarity over telephone calls (Weismer, 2008). In addition, speech pathology research also has a long-established history of studying the intelligibility of disordered speech. A strand of speech pathology research also investigated the articulatory features associated with the loss of speech intelligibility, a similar, but not identical approach to the one taken in SLA (which has focused on individual phonemes and suprasegmental features). As for the field of SLA, intelligibility and comprehensibility tended to be used interchangeably before Munro and Derwing's (1995a) seminal study laid down the theoretical foundation for the separation of accentedness, intelligibility and comprehensibility. Since then, pronunciation research and teaching has entered what Murphy and Baker (2015) termed the fourth wave.

The field's current understanding of the constructs of intelligibility, comprehensibility, and foreign-accentedness stem from Munro and Derwing's (1995a) seminal study, in which they elicited extemporaneous speech from Mandarin-accented L2 English speakers through a picture description task, and presented sections of the recordings to native English listeners who transcribed the sections and assigned Likert-scale ratings of perceived ease of understanding (comprehensibility) and perceived degree of foreign-accentedness. They found that transcription accuracy (intelligibility) was most strongly correlated with perceived ease of understanding (comprehensibility), while it was only weakly correlated with the perceived degree of foreign-accentedness. These results have since been replicated in L2 English using read-aloud sentences (Jułkowska & Cebrian, 2015), and picture-elicited words (Uchihara, 2022). Furthermore, the same separation of the three constructs has been observed in L2 Spanish (Nagle & Huensch, 2020), L2 Mandarin (Neal, 2022), and L2 Arabic (Ali, 2023). In most cases, the strongest correlation was found between comprehensibility and intelligibility, followed by a moderate-to-strong correlation between comprehensibility and accentedness, and either low or no correlation between intelligibility and accentedness, pointing to comprehensibility as the superior predictor of listener understanding of L2 utterances.

In psycholinguistic terms, comprehensibility taps into *processing fluency*, that is, the speed of the online processing of speech (Trofimovich et al., 2022). This theoretical interpretation is exemplified in the empirical results on the connection between comprehensibility and processing time. Munro and Derwing (1995b) presented true and false statements spoken by native and non-native English speakers to native listeners, who were required to assign a truth value to each statement. Response latencies were calculated based on the time listeners took to decide whether the statement was true or false. The response latency to

foreign-accented statements was longer, but this difference stemmed not from higher foreign-accentedness ratings, but from speech that was rated low on comprehensibility. Uchihara (2022) investigated the relationship between processing time and comprehensibility at the word level. In his study, processing time was operationalized as the time elapsed between listening to the word and the first keystroke in transcription. He found that reduced comprehensibility predicted longer processing time more strongly than higher accentedness did.

Reduced speech intelligibility and comprehensibility bear implications on the success of target language interactions. Speech that puts a high processing burden on the interlocutor can not only hinder listening comprehension but also the ability of the interlocutor to successfully participate in and contribute to the interaction. In a series of psycholinguistic experiments, Lev-Ari et al. (2018) discovered that native speakers of English who listened to non-native speech performed more slowly on a lexical recall task and were also less accurate in recalling their own responses to interview questions read by a non-native (Mandarin-accented) researcher. The researchers explained the results as a lower level of detail in general linguistic processing as a result of being exposed to non-native speech. These results could explain the findings of Varonis and Gass (1982), and Gass and Varonis (1985), who uncovered that the reduced comprehensibility of non-native speech causes native speakers to simplify their own speech and sometimes leads them to cut the interaction short altogether. Such reactions from interlocutors could seriously hurt language learners' opportunities for interaction and practice, potentially holding them back from reaching high levels of L2 attainment.

2.2. Segmental Error Gravity in L2 Speech Comprehensibility and Accentedness

L2 pronunciation research has since investigated a multitude of factors involved in speech comprehensibility and accentedness, such as speaker- and listener-based variables, and

linguistic (phonological, lexicogrammatical, pragmatic, and fluency-related) correlates (Crowther et al., 2022). Out of these, the ones that concern L2 pronunciation pedagogy the most are phonological correlates, which could be divided into segmental and suprasegmental ones. Whether it is segmentals or suprasegmentals that are more important and that should take precedence in pronunciation teaching has been an object of a debate (Wang, 2022; Zielinski, 2015) that remains unsettled. The relative importance of these two aspects of pronunciation likely depends on the language under study. Considering the relative scarcity of research on the phonological factors influencing speech comprehensibility in languages other than English, such a debate is unlikely to be fruitful. What has been more successful is finding an organizing principle for an error gravity hierarchy between segments (consonants and vowels) in the form of the Functional Load (FL) principle. The following sections delve into the specifics of the effects of segmental errors on L2 speech comprehensibility and accentedness and chart the development of the studies investigating the predictive power of phoneme FL in segmental error gravity.

2.2.1. Segmental Errors

Segmental errors can be categorized into four types, based on the departure they represent from the native syllable structure of the target word: substitution, distortion, insertion, and deletion (Derwing & Munro, 2015). Segmental substitutions refer to the replacement of the target phoneme with another phoneme of the target language (e.g., pronouncing “think” as “sink” in English, or pronouncing حرام [ħaraːm] as خرام [xaraːm] in Arabic). Distortions are similar to substitutions in that they result in non-target-like production of the phoneme in question, but such production does not involve a recognizable target language phoneme (such as pronouncing the approximant [ɹ] in English as a trill [r] in the case of Arabic-accented English). Insertions and deletions, on the other hand, change the syllable structure of the produced word. While insertion

involves the addition of a phoneme that was not part of the original structure of the word, deletion results in the removal of a phoneme originally present. Most studies on L2 pronunciation have investigated the effects of segmental substitutions, and in the present review, the term “errors” is used to refer to substitutions.

Some studies have compared the differential effect of consonant vs. vowel mispronunciations on L2 speech comprehensibility and accentedness. These have yielded conflicting results. Bent et al. (2007) found that in the case of Mandarin-accented read-aloud sentences in L2 English, vowel mispronunciations harmed intelligibility more than consonantal errors. This is in opposition to the results of Suzukida and Saito (2021) showing that consonants errors were more impactful in the intelligibility ratings of Japanese-accented L2 English extemporaneous speech. A similar result was repeated by Na (2021) on Korean-accented read-aloud English words. In all likelihood, a direct comparison between consonants and vowels might not be useful, since they occupy different positions within the syllable, and possess different FL values, which have language-specific distributions. At the same time, as it will be discussed later, separate hierarchies between different consonants and between different vowels can be determined based on the FL principle.

The position of consonantal errors seems to have an influence on the intelligibility scores of speech segments. Bent et al. (2007) found that Mandarin-accented read-aloud L2 English sentences containing word-initial consonant errors received the lowest intelligibility scores. In their study, this was the only position in which consonant errors were significantly associated with reduced intelligibility. In terms of lexical competition, it makes sense that word-initial mispronunciations would have a more severe impact, since upon hearing the first sound of the word, the listener is sent down the wrong path and it becomes difficult for them to successfully

identify the word after the activation of unrelated competitor words (Mattys et al., 2012). This means that the comparison of consonantal errors needs to take into account the position of said consonants within the word.

Generally, the longer the utterance is, the easier it is to understand it and the same is true for individual words. In Uchihara (2022)'s investigation of word-level intelligibility, comprehensibility and accentedness in L2 English, the number of syllables was a significant predictor of better comprehensibility. This means that words containing fewer syllables were harder to understand. This makes sense when considering that longer words contain more information, and especially in the case of foreign-accented speech, differences from the mental representation of lexical items in L1 listeners can make it particularly difficult to identify the target as the amount of phonological information available decreases. The easier understanding of longer utterances has also been approached from the standpoint of perceptual learning and adaptation. Given enough information about a speaker's phonetic variability, native listeners are able to adapt and learn non-native pronunciation patterns, which has an effect on comprehensibility ratings. However, it is important to point out that this adaptation is subject to individual variation. Nevertheless, when comparing the effects of different consonantal mispronunciations, word length is an important factor to control for.

Errors also seem to have a cumulative effect on comprehensibility and accentedness ratings, although this effect differs based on the FL value of the erroneously produced segment. In addition, the accumulation of errors affects comprehensibility and accentedness ratings differently. Munro and Derwing (2006) found a cumulative effect only for high FL consonant errors on accentedness ratings. This means that the number of consonantal mispronunciations did not affect comprehensibility ratings, nor did the number of low FL consonant errors affect

accentedness. As a replication and extension to the latter study, Alnafisah et al. (2022) included sentences with as many as four consonantal errors, and found that the effect of high FL consonant errors was only magnified when the number of them reached four within a sentence. Low FL consonant errors started showing a cumulative effect earlier (although the effect still remained weaker compared to high FL errors). As for accentedness, high FL errors had a more linear cumulative effect, compared to low FL errors that showed a cutoff after two errors. While these results suggest that the frequency of mispronunciations could be more confusing for listeners (especially in the case of the more important consonants), it is important to point out the myriad of potentially confounding factors that could interfere with the cumulative effects of mispronunciations, including word length, word position within the sentence, error position within the word, as well as the unequal distribution of errors between content words and function words (which Munro and Derwing (2006) highlighted as a limitation in their study).

2.2.2. The Predictive Power of Functional Load in Segmental Error Gravity

By far the most robust predictive framework for segmental error gravity has been the Functional Load (FL) principle. Originally developed within the functionalist circles of the Prague school to explain and predict historical sound change, the classical conceptualization of FL refers to the amount of contrastive work performed by a phonemic opposition (King, 1967): phoneme pairs that differentiate between more minimal pairs in a language have higher FL. Since its introduction into foreign language teaching, the predictive value of the framework has been confirmed by a number of empirical studies. The following paragraphs present a discussion of the historical and theoretical background of the concept, a summary of the empirical evidence, as well as a highlight of gaps and factors that have not been controlled for previously.

The contrastive work of phoneme pairs was traditionally operationalized as the frequency of said phonemic opposition, meaning that phoneme pairs that differentiate more minimal pairs in the lexicon were considered to have higher FL. The original utility of the concept lay in explaining diachronic sound change in language systems: phonemic pairs with higher FL were hypothesized to be more resistant to mergers, since they perform a lot of contrastive work to keep the meanings of lexical items apart (King, 1967). The loss of such a phonemic contrast would potentially hurt communication more than the loss of a low FL contrast. In terms of information theory, the loss of a high FL phonemic contrast would lead to a high level of information loss (entropy) from a particular linguistic system.

Brown (1988) was the first to introduce the concept of FL into foreign language teaching. His formulation of FL could be considered an expansion of King's (1967) definition: it includes 12 considerations, including the part of speech of the minimal pairs, the phonetic similarity of the phoneme pair, and the probability of occurrence of each member of the pair, among others. Arguing for a relative weighting of these 12 factors, he developed a 1–10 ranking of vowel and consonant contrasts in British English. A similar hierarchy was created by Catford (1987) on a scale of 1–100. The main takeaway from Brown's discussion of FL is that FL is more than just the raw cumulative type frequency of the phonemic pair in question. Most importantly, from the standpoint of L2 pronunciation, he argues that only contrasts that are frequently conflated by language learners should be examined.

The hypothetical error gravity hierarchies built by Brown (1988) and Catford (1987) have gained empirical confirmation in L2 pronunciation studies. Munro and Derwing (2006) found that in the case of Cantonese-accented English read-aloud sentences, high FL consonant errors affected both comprehensibility and accentedness ratings more negatively than did low FL

errors. Expanding on these results, Suzukida and Saito (2021) used recordings of extemporaneous speech produced by Japanese-accented L2 English speakers, replicating the negative effect of high FL consonant mispronunciations on comprehensibility ratings. In their study, vowel errors and low FL consonant errors did not have a significant effect on comprehensibility. In another replication of previous findings, Alnafisah et al. (2022) included speakers from multiple language backgrounds in their study. The read-aloud English sentences from these participants containing high FL segmental mispronunciations were judged less comprehensible and more accented than their low FL counterparts. When it comes to vowels, Thir's (2020) study provides the first empirical evidence suggesting that high FL vowel mispronunciations might cause more problems for listeners when compared to low FL ones. The first study conducted on a language other than English found the same negative effect of high FL segmental errors on comprehensibility compared to low FL ones in the case of L2 Chinese speech (Bao et al., 2022). While the robustness of FL in segmental error gravity is increasingly evident from the emergence of methodologically innovative studies, some factors remain to be controlled for, such as segmental error position within the word, grammatical category, and even word length.

2.3. The Relationship Between Functional Load and Consonant Age of Acquisition

Owing to FL's robustness as a predictive framework for segmental error gravity, Munro and Derwing (2015) call for the exploration of FL hierarchies and their effects on speech comprehensibility and accentedness in languages beside English. This call has been echoed by researchers in the field of AFL (Hellmuth, 2014; Rifaat, 2017; Wahba, 2021). Hellmuth (2014) points out the lack of a clearly identified FL hierarchy in Arabic. The difficulty of identifying such a hierarchy is exacerbated by a dearth of representative, phonetically annotated spoken

corpora (Ahmed et al., 2022). Relying on early evidence on the effect of FL on L1 acquisition (Stokes & Surendran, 2005) in certain languages, Hellmuth proposes an error gravity hierarchy based on the order of acquisition of consonants as described by Amayreh and Dyson (1998) for Jordanian children. However, such proposals need to gain empirical confirmation before implemented pedagogically. Therefore, the aim of this section is to outline the empirical and theoretical underpinnings that establish the relationship between FL and consonant age of acquisition (CAoA) and to discuss the hypothesized relationship between FL and CAoA in Arabic. The section then concludes with a review of L1 phonological acquisition studies in Arabic and provides a categorization the most common consonantal errors L2 speakers of Arabic make based on the CAoA data available in Egyptian Arabic.

The study of first language acquisition, similarly to that of SLA, has traditionally been dominated by formalist explanations for acquisitional patterns. These employ linguistic universals such as markedness and articulatory complexity to explain why certain phonemes are acquired earlier than others. According to this explanation, phonemes that are more marked are acquired later than unmarked ones: e.g., voicing is a marked phonological feature and as such, voiced consonants (e.g., /d/) are acquired later than their unmarked counterparts (e.g., /t/). In terms of articulatory complexity, fricatives are more complex to produce articulatorily, and therefore are acquired later than stop consonants. In addition, the traditional generativist view (itself being a formalist approach) views the acquisition of language as top-down, with an innate language acquisition device that predisposes children to follow similar paths of acquisition across languages. The dominance of linguistic universals, however, has come under question by proponents of the functionalist usage-based (emergentist) approach (Diessel, 2017). Usage-based or emergentist approaches to language acquisition seek to locate the emergence of linguistic

forms in children's usage within the ambient language that the child is exposed to. According to the usage-based framework, no top-down innate device is needed, but rather the child makes generalizations from bottom-up observations (Behrens, 2009). Based on this framework, the language-specific acquisitional patterns that arise can be explained by characteristics of the linguistic input that the child comes in contact with.

In the case of phonology, the two most common language-specific predictors of acquisition order are phoneme frequency and phoneme FL (Ingram, 2008; Tribushinina & Gillis, 2017). While the traditional calculation of FL involves determining values for specific phoneme pairs, the computational, information-theoretical approach allows for the computation of FL values for individual phonemes by pairing them with articulatorily similar counterparts and arriving at an approximate value. Sewell (2017) terms this formulation the *broad sense* of FL, and it has been found to be more robust than Brown's (1988) hierarchy in predicting CAoA (Severen et al., 2013). In addition, the computational formula of FL takes into consideration token frequency, since it is used to calculate segmental FL within representative spoken corpora.

FL has been found to be as or more predictive than phoneme frequency for CAoA. Stokes and Surendran (2005) found that FL was a unique predictor of CAoA in English and Dutch, where they did not find additional explanatory power for frequency. In the case of Cantonese, only frequency seemed to be a predictive factor in CAoA. In order to make more valid assumptions about children's ambient language influences, Severen et al. (2013) used child-directed speech corpora and found that the FL calculations based on them better predicted CAoA for Dutch word-initial consonants compared to FL calculations based on adult-directed speech corpora. They also found that token frequency was a better predictor than type frequency, which does not take into account non-standard pronunciations. In a larger-scale comparison of five

languages (English, Japanese, Mandarin, Turkish, Spanish), Cychosz (2017) found that FL (calculated based on child-directed speech) was a better predictor of CAoA than frequency in four of them, with only Mandarin showing a reverse pattern. Overall, these results point to likely effects of language typology, as tonal languages do not seem to show high FL effects on CAoA. In all likelihood, the more a language relies on vowels and suprasegmentals to contrast meaning, the less role its consonant FL distribution plays in consonant acquisition order.

2.3.1. The Hypothesized Relationship Between Functional Load and Consonant Age of Acquisition in Arabic

Froud and Khamis-Dakwar's (2021) critical review of Arabic L1 acquisition studies highlights that most of the published literature on Arabic L1 phonological acquisition has approached the subject from the standpoint of universal processes. The discussion of FL as a possible explanation for Arabic-specific acquisitional patterns appears in Amayreh and Dyson (1998), Amayreh and Dyson (2000), and Amayreh (2003). However, they do not quantify FL in any of their studies, merely suggesting that those consonants that defy cross-linguistic patterns of acquisition based on their marked or articulatorily complex features could have higher FL values in Arabic: e.g., the voiceless pharyngeal fricative /ħ/. They are also overly cautious in highlighting that many late acquisition consonants seem to also be marked or articulatorily complex (e.g., the emphatic consonants with pharyngeal secondary articulations). This interpretation seems to be an unnecessary privileging of linguistic universals over language-specific explanations. As it will be discussed, FL could still correlate with CAoA even in cases of overlap between predictions made by linguistic universals.

By looking at the phonological structure of the different languages in Cychosz's (2017) study, we could hypothesize the likely magnitude of FL effects on Arabic CAoA. In her

investigation, Spanish showed the strongest correlation between consonant FL and CAoA, a relationship that is stronger than the one observed in English. There are indications that this relationship in the case of Arabic could be stronger than the one in English, possibly approaching the one found in Spanish. Firstly, unlike English, Arabic does not have contrastive stress. English listeners seem to be sensitive to stress errors, which affect speech comprehensibility negatively. Stress errors in English also introduce vowel errors, since unstressed vowels undergo reduction, a phenomenon that is similarly absent from Arabic. In terms of its vowel inventory, Arabic has three to five distinct vowel qualities (as opposed to 25-28 consonant phonemes depending on the dialect), which is fewer than the five vowels found in Spanish, and much fewer than the 10+ vowels that exist across varieties of English. What could put Arabic behind Spanish is its two contrastive suprasegmental features: vowel length and consonant gemination. In comparison, Spanish does not have contrastive suprasegmental features, relying only on consonants and a small set of vowels to differentiate meaning. In this sense, Arabic could potentially exhibit an FL-CAoA relationship that is between Spanish and English, which would mean a fairly high correlation.

These typological differences between the different languages under study are represented in the language-specific patterns of early spoken word recognition by children. While infants from multiple language backgrounds exhibit a vowel bias in spoken word processing in the first year of age, this bias shifts in favor of consonants in languages that make greater use of consonantal contrasts (Nazzi & Cutler, 2019). For example, this shift takes place by the 12th month in Spanish-speaking children (Bouchon et al., 2022). As presented before, Spanish displays a very strong correlation between consonant FL and CAoA. In comparison, English-learning children only develop a consonant bias within their 3rd year of life (Ratnage et

al., 2023), which can explain why English shows a weaker correlation between FL and CAoA than Spanish. Lastly, Cantonese- and Mandarin-speaking children retain the vowel advantage even in their 3rd year of age (Chen et al., 2021; Ma et al., 2017). In addition, speakers of the latter two languages still do not display a consonant bias in speech processing even in adulthood. This again can explain the weak correlation between consonant FL and CAoA in both Cantonese and Mandarin.

When it comes to the consonant bias in Arabic spoken word processing, psycholinguistic evidence favors a strong consonant advantage. Aldholmi and Pycha (2023) conducted two experiments to investigate the differential effects of the removal of vowels vs. consonants from MSA stimuli and found that sentences that had their vowels masked were more accurately identified by listeners than sentences where the consonants were masked and only the vowels could be heard. This finding concurs with similar results found in other Semitic languages like Hebrew (Lador-Weizman & Deutsch, 2022). In comparison, sentence-level word recognition by English-speaking adults shows a vowel bias. The finding also makes sense in light of the typological distribution of segments in Arabic, where the balance of the scale is tipped towards consonants. However, despite claims to the contrary, this consonantal bias seems to be further enhanced by the root-and-pattern-based morphology of Semitic languages. Lador-Weizmann and Deutsch (2022) compared the consonant bias for morphologically complex and morphologically simple Hebrew words and found that it was stronger for the complex words that had clearly identifiable Hebrew roots and patterns.

Computational evidence also points to the advantage that consonants offer to children learning a root-and-pattern-based language like Arabic. Kaastner and Aadrians (2018) used a Bayesian computational model that approximates the statistical learning that children are thought

to engage in when processing the ambient language, in accordance with usage-based approaches to language acquisition. They compared the performance of this model on Arabic and English phoneme segmentation by feeding it both consonant-only data and consonant-and-vowel (full) representations. In the case of Arabic, the model was more accurate at correctly identifying word and morpheme boundaries when fed consonant-only data than when it was given both consonants and vowels. In English, this consonant advantage was not present, suggesting that a young learner of Arabic benefits from ignoring vowels and focusing on consonants to acquire the language.

Overall, the phonological structure that favors consonants against vowels and suprasegmentals, the unique morphological utilization of consonants in the form of roots, and psycholinguistic evidence for a consonant bias in speech processing all point to the possibility that CAoA is strongly associated with FL in Arabic. That is, early-acquisition consonants likely represent high FL values and late-acquisition consonants likely represent low FL in the language. Thus, the hypothesis put forth by Hellmuth (2014) seems plausible. What is needed, now, is a more thorough examination of CAoA in Arabic, with the recognition that dialects probably differ in the order of acquisition of their consonants, likely pointing to differing FL distributions. This does not come as a surprise, considering that FL is a usage-based concept that is supposed to reflect the changing nature of language and social and regional variation in patterns of language use.

2.3.2. Arabic L1 Consonant Acquisition Order

While many studies on Arabic phonological acquisition suffer from restrictiveness in scope (small sample sizes, limited age ranges, non-standard elicitation methods, different criteria for acquisition; Froud & Khamis-Dakwar, 2021), there have recently been promising, large-

scale, cross-sectional investigations of CAoA in Syrian (160 children between the ages of 2:6–6:5; Owaida, 2015) and Egyptian Arabic (360 children between the ages of 1:6–7:4; Elrefaie et al., 2021). The results of these can be compared with Amayreh and Dyson's (1998) large-scale investigation of CAoA in Jordanian Arabic (180 children between the ages of 2:0–6:4). According to Froud and Khamis-Dakwar (2021), differences observed could be indicative of dialect-specific acquisitional patterns. For this reason, a speech perception study relying on CAoA data needs to include listeners from a single dialectal background, to reliably infer underlying FL values. Listeners also bring their unique dialectal experiences that could affect their comprehensibility ratings (e.g., the lexicon of each dialect is different). This suggestion is also in accordance with Sewell's (2017) note on FL being a context-dependent, *dynamic* phenomenon, as opposed to being a fixed property of a larger, more abstract linguistic structure.

L1 consonant acquisition has traditionally been categorized into three stages: early, middle, and late acquisition. However, these are merely relative stages, since studies have differed in the age ranges of the children they included as participants across languages. For example, McLeod and Crowe (2018) reviewed studies describing English consonant acquisition and categorized early as comprising the age range 2;0–3;11, middle as 4;0–4;11, and late as 5;0–6;11. However, for Korean, the early stage spans 2;0–2;11, the middle stage spans 3;0–3;11, and the late stage spans 4;0–4;11. In the case of Arabic, Amayreh and Dyson (1998) categorized their data as early (2;0–3;10), intermediate (4;0–6;4), and late (after 6;4). From the standpoint of segmental error gravity, it is more practical to divide the stages of acquisition into early- and late-acquisition consonants, in keeping with the studies comparing segmental errors with high and low FL. The boundary between them can be drawn around age 4, which seems to be the midpoint of acquisition across the three large-scale studies conducted on Arabic. It is also the age

by which the majority of consonants have been shown to acquire in the 27 languages reviewed by McLeod and Crowe (2018).

Importantly, we need to remember Brown's (1988) point about FL: we are only interested in the error gravity hierarchies of phonemes that are likely to be mispronounced by L2 learners of the language. Fortunately, we can rely on Al Tubuly (2018)'s descriptive investigation of the production accuracy of Arabic consonants by L2 learners. This study involved 50 L2 learners of Arabic from five language backgrounds: English, German, Greek, Turkish, Chinese. The advantage of this investigation is that mispronunciations originated from native speakers of phonologically diverse languages, giving a more complete account of the likely difficulties L2 learners of Arabic face when pronouncing segments. The consonants that exhibited lower than 90 percent accuracy in the study in production were the following: /h/ (ه), /x/ (خ), /ð/ (ذ), /ɣ/ (غ), /sˤ/ (ص), /dˤ/ (ض), /ðˤ/ (ظ), /tˤ/ (ط), /ʕ/ (ع), /q/ (ق), /ħ/ (ح). Out of these, the voiced interdental fricative /dh/ (ذ) is absent from most colloquial dialects, including Egyptian and Syrian Arabic, and native speakers of these dialects usually substitute it with the voiced alveolar fricative /z/ (ز). Additionally, while the glottal stop /ʔ/ (ء) showed 100 percent accuracy in Al Tubuly's study, its substitution with the voiced pharyngeal fricative /ʕ/ (ع) seems to be common by L2 learners and can oftentimes be observed in word-initial position (e.g., pronouncing أَمْر /ʔamr/ as عَمْر [ʕamr]. This leaves us with 11 consonants most likely to be mispronounced by L2 learners of Arabic for which establishing an error gravity hierarchy would be the most useful: /h/ (ه), /x/ (خ), /ɣ/ (غ), /sˤ/ (ص), /dˤ/ (ض), /ðˤ/ (ظ), /tˤ/ (ط), /ʕ/ (ع), /q/ (ق), /ħ/ (ح), /ʔ/ (ء).

There tend to be terminological discrepancies between studies with regards to the usage of acquisitional criteria. In the case of the three large-scale Arabic studies, *acquisition* was used to mean 75 percent of children in an age group correctly producing the consonant in each

position tested in the case of Jordanian Arabic (Amayreh & Dyson, 1998), while it referred to a 90-percent criterion in the case of Syrian (Owaida, 2015) and Egyptian (Elrefaie et al., 2021). The 90-percent criterion was called *mastery* in the Jordanian study, while mastery was used to denote a 100-percent criterion in the Egyptian study. Due to these discrepancies, the shared 90-percent acquisitional criterion can be used to categorize the potential mispronunciations into early- and late-acquisition groups. Interestingly, using this criterion, all 11 consonants would fall under late acquisition in the case of Jordanian Arabic. In fact, another study by Amayreh (2003) found that by age 8;4, there were still consonants unacquired by Jordanian children. These unusually late acquisitional results go against both the more recent evidence on Syrian and Egyptian Arabic, where children were found to acquire most consonants by age 6 (except /q/, /ʒ/, and /ðˤ/ in Syrian), and the crosslinguistic evidence showing that almost all consonants are acquired by age 6 (McLeod & Crowe, 2018).

Following the 90% criterion and the 4-year boundary between early and late acquisition, both Syrian and Egyptian data point to the same categorization of the 11 consonants of interest (Elrefaie et al., 2021; Owaida, 2015). Early-acquisition consonants are the following: /h/ (ه), /ʕ/ (ع), /ħ/ (ح), /ʔ/ (ء); while late-acquisition consonants include the following: /x/ (خ/, /ɣ/ (غ), /tˤ/ (ط), /dˤ/ (ض), /sˤ/ (ص), /ðˤ/ (ظ), /q/ (ق). Of course, data from other dialects would be useful to enrich this categorization or to potentially develop other categorizations based on dialect-specific data. However, this is not possible because studies have been conducted on specific, restricted age groups and thus are not directly comparable with each other.

2.4. Conclusion

Overall, CAoA seems to be a promising predictor variable in consonantal error gravity in languages where the contrastive distribution of segments and suprasegmentals favors consonants.

In such languages, CAoA shows a significant and strong correlation with FL. Drawing on languages with similar phonological paradigms, it can be hypothesized that CAoA is strongly associated with FL in Arabic. Thus, the expectation is that early-acquisition consonants represent high FL values, while late-acquisition consonants represent low FL values. Consequently, early-acquisition consonant errors should reduce speech comprehensibility more than late-acquisition errors. Likewise, early-acquisition consonant mispronunciations should increase the perceived degree of foreign-accentedness more than late-acquisition ones. Overall, indirect inferences about underlying FL distributions can be of great use in the case of other under-researched languages with few available and representative spoken corpora.

Chapter Three: Methodology

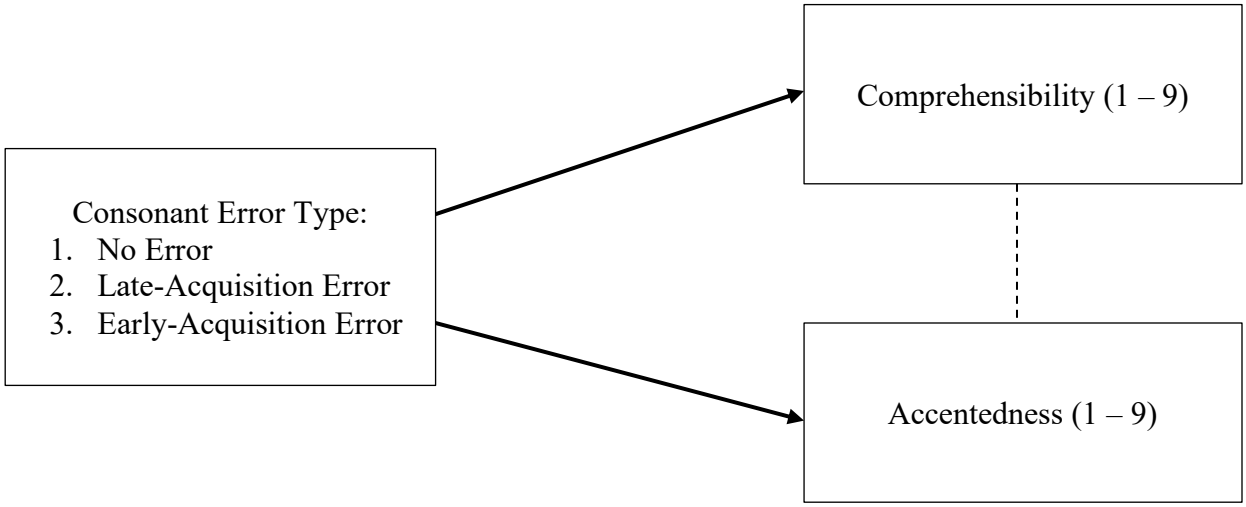
This chapter describes the methodology of the current investigation. It starts with a description and justification of the operational variables and the study design, followed by an outline of the research procedure from beginning to end. In subsequent sections, each step in the procedure is discussed in detail to present the rationale for the methodological choices. Participants come from two populations: L2 learners of Arabic who acted as speakers and NSs of Egyptian Colloquial Arabic (ECA) who acted as listeners in this study. Accordingly, both the speaking task and the listening task are discussed in detail in their separate sections. Additionally, since both sets of participants filled out Language Background Questionnaires in addition to their individual tasks, this is also discussed in its separate section. Finally, data analysis decisions are described and justified.

3.1. Study Design and Variables

The current investigation is a quantitative speech perception study aimed at exploring the relationship between the independent variable of error type and two dependent variables: comprehensibility, and accentedness. The design of the study can be seen in Figure 3.1. Error type is a categorical, within-subjects variable consisting of three levels: early-acquisition error, late-acquisition error, and no error. This is in keeping with the original operationalization of error type by Munro and Derwing (2006), who compared sentences containing no errors, sentences containing low FL errors and sentences containing high FL errors. Comprehensibility and

Figure 3.1

Study Design



accentedness are operationalized as 9-point Likert scales, with higher numbers indicating more difficulty in understanding, and a higher degree of perceived foreign accent, respectively. As discussed in the literature review, these two dimensions of speech are well-established in L2 pronunciation research and can only be measured through subjective, numerical listener ratings. Since spoken interactions occur between speakers and listeners in real life, breakdowns in communication can only be predicted through listener reactions to L2 speech. In addition to examining the effect of the independent variable on the dependent variables, the strength of the relationship between the two dependent variables was also of interest, to further add to the research examining the interrelationships between dimensions of speech for different task types. Since this is the first study using read-aloud words to elicit ratings for both comprehensibility and accentedness, it is important to measure the correlation between the two for this task.

3.2. Procedure

This section summarizes the steps in the research procedure. The study largely followed the standard procedure used in L2 pronunciation studies, as described by Munro and Derwing

(2015). First, the speaking task was designed (read-aloud word list) by selecting words from an Arabic frequency dictionary. After receiving IRB approval (see Appendix A), speakers (L2 learners of Arabic) were recruited to record the word list (see Appendix B) using their own electronic devices (mobile phones or computers). They signed an informed consent form (see Appendix C) and filled out a Language Background Questionnaire (see Appendix D) to provide descriptive information about their linguistic background. Two NSs of Egyptian Arabic were also recruited to record the same word list. After receiving the recordings, they were edited using Audacity. Recordings of individual words were then selected to represent each type of consonantal mispronunciation. Also, control words from L2 speakers not containing any mispronunciation were selected, as well as recordings from the native speakers as another control group. Then, the listener survey was designed using Qualtrics and it was piloted to ensure clarity of instructions and usability of the interface. Following the pilot, the survey was distributed among participants using an anonymous link, who indicated their informed consent (see Appendix E for consent form), rated the recordings on comprehensibility and accentedness (see Appendix F for general task instructions and Appendix G for the rating scales), and filled out a Language Background Questionnaire (see Appendix H). This was followed up with data analysis by computing descriptive statistics, interrater reliability and conducting inferential statistical tests. The following sections detail each of these steps and discuss the rationale for the methodological choices.

3.3. Stimuli

3.3.1. Speakers

Ten L2 learners of Arabic (six men and four women) studying Arabic at The American University in Cairo (AUC) were recruited to record a list of 40 MSA words. This was similar to

Uchihara's (2022) study, in which 40 individual words were elicited from 12 speakers. Generally, L2 pronunciation studies involve ten to twenty speakers, which is adequate to collect a range of mispronunciations. Each speaker was sent a randomized word list containing the same 40 words, a consent form and a Language Background Questionnaire via email. Five of them were L1 speakers of English, two of them were L1 Polish speakers, one of them was an L1 Chinese speaker, one was an L1 speaker of Italian, and one was an L1 German speaker. All of them indicated that they had an Advanced or Superior level of proficiency in Arabic. Since most of the research done on AFL has involved the participation of L1 English learners of Arabic, it was important to include speakers from multiple L1 backgrounds. In the case of L2 phonology, such diverse sampling is especially important since many errors are influenced by L1, and the probability and range of mispronunciations can be increased by the inclusion of learners speaking a variety of L1s.

3.3.2. Word List

The current study involves a read-aloud speaking task, in which speakers were required to record a list of 40 Modern Standard Arabic words (see Appendix B for word list), divided into two groups: 20 words containing a word-initial early-acquisition consonant, and 20 words containing a word-initial late-acquisition consonant. Table 3.1 contains the dictionary forms of the words, as well as the actual pronunciation of them as they appeared in the study. After receiving feedback from the two native Egyptian speakers who recorded the word list, five words were removed from the analyses, because they contained short vowels that could be considered non-targetlike by Egyptian listeners. After consulting *A dictionary of Egyptian Arabic: Arabic-English* (Badawi & Hinds, 1986), it was confirmed that three of these words have different vowel realizations in Egyptian Arabic (MSA: حُزْمة → ECA: حِزْمة, MSA: حَفْنة → ECA: حِفْنة, MSA:

Table 3.1

List of Experimental Stimuli with Dictionary Forms vs. Actual Phonetic Realization by Speakers

No.	Experimental Condition	MSA Form[1]	ECA Form[2]	Phonetic Realization	Speaker
1	Early-Acquisition	هِبة /hiba/	same as MSA	حِبة [ħiba]	NNS10
2*	Early-Acquisition	هاجِس /ha:dʒis/	هاجِس /ha:gis/	حاجِس [ħa:gis]	NNS1
3	Early-Acquisition	هُدْنة /hudna/	same as MSA	حُدْنة [ħudna]	NNS2
4*	Early-Acquisition	حُزْمة /ħuzma/	حِزْمة /ħizma/	هُزْمة [huzma]	NNS4
5	Early-Acquisition	عَتْمة /ʕatma/	same as MSA	أَتْمة [ʔatma]	NNS4
6	Early-Acquisition	عِنان /ʕina:n/	N/A	إنان [ʔina:n]	NNS4
7	Early-Acquisition	عِداد /ʕida:d/	N/A	إداد [ʔida:d]	NNS4
8*	Early-Acquisition	أَمَد /ʔamad/	N/A	عَمَد [ʕamad]	NNS7
9	Early-Acquisition	هَيْكَل /hajkal/	same as MSA	حَيْكَل [ħajkal]	NNS7
10	Early-Acquisition	هَيْبة /hajba/	هَيبة /he:ba/	حَيْبة [ħajba]	NNS7
11	Late-Acquisition	ضَجيج /dˤadʒi:dʒ/	ضَجيج /dˤagi:g/	دَجيج [dagi:g]	NNS2
12*	Late-Acquisition	صَليب /sˤali:b/	same as MSA	صَليب/سَليب [sałi:b]	NNS2
13*	Late-Acquisition	صَواب /sˤawa:b/	same as MSA	سَواب [sawa:b]	NNS2
14	Late-Acquisition	صارِم /sˤa:rim/	N/A	سارِم [sa:rim]	NNS6
15*	Late-Acquisition	غَمْرة /ɣamra/	N/A	خَمْرة [xamra]	NNS8
16	Late-Acquisition	غارة /ɣa:ra/	same as MSA	خارة [xa:ra]	NNS8
17	Late-Acquisition	غَرام /ɣara:m/	same as MSA	خَرام [xara:m]	NNS8
18	Late-Acquisition	غَفْلة /ɣafla/	same as MSA	خَفْلة [xafla]	NNS8
19	Late-Acquisition	ضَباب /dˤaba:b/	same as MSA	دَباب [daba:b]	NNS8
20	Late-Acquisition	خُلود /xulu:d/	N/A	غُلود [ɣulu:d]	NNS9
21	No Error	حَسْرة /ħasra/	same as MSA	حَسرة [ħasra]	NNS10
22	No Error	حَميد /ħami:d/	same as MSA	حَميد [ħami:d]	NNS1
23*	No Error	حَفْنة /ħafna/	حِفْنة [ħifna]	حَفْنة [ħafna]	NNS2
24	No Error	قَبيل /qabi:l/	N/A	قَبيل [qabi:l]	NNS3
25	No Error	حافِز /ħa:fiz/	same as MSA	حافِز [ħa:fiz]	NNS4
26	No Error	قامة /qa:ma/	قامة /ʔa:ma/	قامة [qa:ma]	NNS5
27	No Error	صَميم /sˤami:m/	same as MSA	صَميم [sˤami:m]	NNS6
28	No Error	أُسْوة /ʔuswa/	N/A	أُسْوة [ʔuswa]	NNS7
29	No Error	خانة /xa:na/	same as MSA	خانة [xa:na]	NNS8
30	No Error	خَيْبة /xajba/	خَيبة [xe:ba]	خَيْبة [xajba]	NNS9

*: excluded from all analyses

[1]: Buckwalter and Parkinson (2011)

[2]: Badawi and Hinds (1986)

عَنْوة → ECA: عِنْوة). One of them did not appear in the Egyptian dictionary (pronunciation in

Table 3.1 continued

List of Experimental Stimuli with Dictionary Forms vs. Actual Phonetic Realization by Speakers

No.	Experimental Condition	MSA Form	ECA Form	Phonetic Realization	Speaker
31**	Native Speaker	ظُلْمة /ðˤulma/	ضَلْمة /dˤalma/	ظُلْمة [ðˤulma]	NS1
32**	Native Speaker	خَلَف /xalaf/	same as MSA	خَلَف [xalaf]	NS1
33**	Native Speaker	قالِب /qa:lib/	قالِب /ʔa:lib/	قالِب [qa:lib]	NS1
34*	Native Speaker	عَنْوة /ʕanwa/	عِنْوة /ʕinwa/	عَنْوة [ʕanwa]	NS1
35*	Native Speaker	حَلْبة /ħalba/	same as MSA	حَلَبة [ħalaba]	NS1
36**	Native Speaker	خَراب /xara:b/	same as MSA	خَراب [xara:b]	NS2
37**	Native Speaker	حُرْمة /ħurma/	same as MSA	حُرمة [ħurma]	NS2
38*	Native Speaker	حَراك /ħara:k/	N/A	حِراك [ħira:k]	NS2
39**	Native Speaker	قُدوم /qudu:m/	قُدوم /ʔudu:m/	قُدوم [qudu:m]	NS2
40**	Native Speaker	عِبْرة /ʕibra/	same as MSA	عِبْرة [ʕibra]	NS2

*: excluded from all analyses

**: excluded from inferential analyses

study: حِراك, MSA dictionary form: حَراك), and the fifth one did not appear in any dictionary (pronunciation in study: حَلَبة, dictionary form: حَلْبة). For the sake of comparison between MSA and ECA equivalents of words, the table includes the forms of each word from both dictionaries. Upon review, one word was removed from the final analyses because its phonetic realization was ambiguous (سَليب/صَليب). Four additional words were removed, because while they contained substitutions compared to their written form, the consonant substitutions resulted in real words (أَمَد → عَمَد; غَمْرة → خَمْرة; هاجس →حاجس, too similar to حاجز, because final consonant devoicing occurs frequently in ECA; صَواب → سَواب, phonetically identical to ECA pronunciation of ثَواب). Therefore, the final inferential analyses included ratings for 23 words: seven early-acquisition items, seven late-acquisition items, and nine correctly pronounced words. The words produced by natives were not included in the inferential statistical tests and were only used to conduct reliability analyses.

While read-aloud word lists do not represent the most common domain for target language speech of L2 learners, such an approach affords us the necessary experimental control to compare the effects of individual consonants on L2 Arabic speech comprehensibility and accentedness. Importantly, no claims are made about the relative importance of consonants compared to vowels, suprasegmental features and other variables. This means that only comparisons between consonants are made, which can be done on the basis of read-aloud words where other variables are controlled for. Caspers and Horłoza (2012) found no difference between the intelligibility and accentedness ratings of words read aloud from a list vs. words that were extracted from a longer read-aloud text. Using a word list also has the advantage of more efficiently eliciting different types of consonantal errors, compared to waiting for these errors to occur in extemporaneous speech, where speakers could also use avoidance strategies to mask their pronunciation difficulties. In addition, a word-level investigation effectively controls for a myriad of potential confounding variables that could affect comprehensibility and accentedness ratings, such as lexicogrammatical factors (Appel et al., 2019) and speech fluency (Derwing et al., 2004; Munro & Derwing, 2001).

The selection of words for the speaking task took into consideration the following factors: a) frequency; b) grammatical category; c) length; d) vowel structure; and e) the position of the potentially mispronounced consonant. In order to maximize the likelihood of speakers committing pronunciation errors, medium-frequency MSA lexical items were selected from *A frequency dictionary of Arabic: Core vocabulary for learners* (Buckwalter & Parkinson, 2011). Since no systematic frequency data are available on colloquial lexical items, it was necessary to use MSA items. Lexical items are considered medium frequency between the ranks of 3000 and 9000 (Vilkaitė-Lozdienė & Schmitt, 2019). Although Uchihara used low-frequency (2022)

words in the case of English, this was not possible in Arabic due to the unavailability of systematic data beyond the 5000 frequency level in the dictionary. Generally, the goal was maximizing the number of potentially unknown words to the speakers, possibly exposing them to these lexical items for the first time as they encounter them on the list, which could increase the chances of them not having mental phonological representations for these words. A secondary objective of using medium-frequency words was to minimize the effects of word familiarity on listeners' ratings of comprehensibility.

Additionally, the grammatical category of the words was made constant with the sole selection of content words: mostly nouns and some adjectives. Munro and Derwing (2006) pointed to the potential confounding effect of grammatical category (content words vs. function words) in their study, and this aspect has not been controlled for in subsequent studies on FL effects. An attempt was made to use underived nouns (أسماء جامدة) and adjectives, to ensure that the words start with their first root consonant. An exception was made for verbal nouns of verbs from pattern I because their morphological patterns do not include additional non-root consonants. This is to mitigate any potential confounding effects of the unique root-pattern morphology of Arabic on listeners (Gwilliams & Marantz, 2015). Furthermore, it was believed that the decontextualized presentation of nouns and adjectives would be more natural than that of verbs.

As for the phonological structure of the words, the word list only included disyllabic items containing no final consonant clusters and only one long vowel. This was to ensure a) that word length does not confound the comprehensibility results (Uchihara, 2022) and b) that only consonant errors take place. Since vowel length difference tends to be difficult for many L2 learners of Arabic, the inclusion of words with two long vowels (e.g., قاموس) was avoided.

Similarly to vowel length, geminated (i.e., doubled or long) consonants (ː) were avoided, since many of the students' L1s lack this feature, and geminated consonants tend to be difficult to produce. The difficulty of final clusters also could have potentially caused speakers to insert helping vowels between them, introducing departures from the original word structure beyond consonant substitutions.

Finally, the words were selected in a way to include the consonants potentially difficult to L2 speakers in word-initial position only. This is following evidence provided by Bent et al. (2007) for the gravity of word-initial consonant errors compared to errors in other positions. In addition, none of the words contained more than one potentially difficult consonant, in keeping with the findings on the effects of multiple errors on comprehensibility and accentedness ratings (Alnafisah et al., 2022; Munro & Derwing, 2006). In general, the words were selected in a way to maximize the occurrence of one and only one consonant error, in word-initial position. Overall, the list was divided into two equal parts: 20 words starting with difficult early-acquisition consonants, and 20 words starting with difficult late-acquisition consonants.

3.3.3. Recording

The speakers conducted their recordings on their personal electronic devices, in the comfort of their homes or other, quiet public places. They were instructed to not rehearse the word lists in any way, but to simply start recording whenever they are ready and are in a quiet place and read the list of 40 words in one go, leaving a one-second silence in between them. Participants submitted their recordings via email. Crowther and Urada (2022) compared sound recordings elicited face-to-face and remotely and found no difference in the sound clarity perceived by listeners. This suggests that remote elicitation of the recordings was a viable option. Following the submission of the recordings by speakers, they were edited using Audacity. First,

the volume of the recordings was normalized since they were not at a uniform level of loudness. Then, noise reduction was performed by filtering out street noise and other background noise. Finally, each recording was split into 40 different ones, each of them containing one word. The words for the survey were selected from these 480 individual recordings coming from the ten L2 speakers and the two NSs, in a way that no lexical item appeared in the survey more than once.

3.4. Listening Task

3.4.1. Listeners

Twenty-eight native speakers of Egyptian Arabic (16 male and 12 female) volunteered to fill out an online survey on Qualtrics. Upon inspecting the responses of participants to the Language Background questionnaire, six of them were excluded due to having indicated language teaching experience, which has been found to affect the way listeners evaluate non-native speech (Saito, 2021). All description and analyses were conducted on the remaining 21 participants (14 men and seven women; mean age=26.9 years, range 19-38). No participant reported experience with linguistic research. Fifteen of them reported no phonetic experience, while six of them responded yes, possibly having training in tajweed (Qur'anic recitation), as it was mentioned as an example in the question. Eighteen of them reported having experience with non-native Arabic speech, and only three of them indicated that they did not. All participants reported knowledge of English as a foreign language, three of them indicated knowledge of German, two of them French, and one participant reported knowledge of Spanish and Portuguese. Overall, four of them were trilingual, and one participant was quadrilingual. In terms of schooling, 15 participants had a bachelor's degree, four participants were master's degree-holders, and two of them had a high school diploma as their highest level of education. Nine of

them indicated receiving their education in English only, six of them received it in Arabic only, and six of them received both Arabic and English education.

3.4.2. Survey

The survey was hosted on Qualtrics, which has been successfully used in previous L2 pronunciation studies. Alnafisah et al. (2022) used Qualtrics to investigate the effects of FL on comprehensibility and accentedness. They obtained highly reliable results, with intraclass correlation coefficients above 0.9 for both comprehensibility and accentedness. Nevertheless, a pilot study was conducted to gather feedback about the viability of the online interface, as well as the clarity of instructions. Four native speakers of Egyptian Arabic completed the survey and none of them reported any technological issues, nor any difficulties in understanding and completing the task. While one participant suggested reducing the number of scale options, this suggestion was not implemented, for reasons that will be discussed in the next section. Since no modifications were made to the survey after the pilot, the responses of these participants were included in the final analyses.

In the current study, the mean time for survey completion was 13.63 minutes (range 8.53-24.73). The survey started with the consent form in Arabic, followed by the general task instructions. Participants were given a simple explanation of the task, along with a notice asking them to wear headphones. They were then presented with two practice ratings: the first was produced by one of the native speakers, and the second one was produced by an L2 speaker and contained multiple mispronunciations (أُسْوة → أَسَوّة). These two words were selected so that listeners gained a sense of the task and practiced using the two ends of the scales. Following this short practice block, a block containing the 40 target words was presented in a randomized order. Each page contained only one recording, alongside the two scales for rating. After this block, the

Language Background Questionnaire was presented, followed by the final page thanking listeners for their participation.

3.4.3. Rating Scales

The survey included 9-point Likert-scales to elicit assessments of comprehensibility and accentedness, as customary in L2 pronunciation research. Trofimovich et al. (2022) summarizes the methodological alternatives that have been explored in L2 comprehensibility research. They concluded that there has been little difference observed between scale types. 9-point scales seem to yield similar results to continuous sliders and also to direct magnitude estimation, where listeners rate every recording in relation to a reference recording. In addition, Uchihara's (2022) investigation of word-level comprehensibility and accentedness used 9-point scales, and found high interrater reliability in the case of both (Cronbach's alpha above 0.9). For the sake of convenience, and comparability of results with the myriad of L2 pronunciation studies employing the same method, the 9-point scale was the most sensible choice. To determine interrater reliability, a two-way, random effects, consistency intraclass correlation coefficient (ICC) was computed for listener's comprehensibility ratings, which indicated excellent interrater agreement, .966, $F(29,580) = 29.81$, $p < .001$, [95% *CI* .946 – .982]. Similarly, the listeners showed excellent interrater agreement for accentedness ratings, .967, $F(29,580) = 29.87$, $p < .001$, [95% *CI* .946 – .982]. Following this, ratings of each listener for the error conditions were averaged for the inferential analyses of the effects of the independent variable.

3.5. Language Background Questionnaire

Munro and Derwing (2015) recommend collecting demographic and linguistic information about participants through Language Background Questionnaires (LBQs). For the speaker LBQ, see Appendix C and for the listener LBQ, see Appendix D. The purpose of the

questionnaire is to uncover individual factors that could potentially affect the results of studies. In the case of the current study, speaker and listener LBQs are used to gather descriptive statistics of the linguistic experience of participants, since such experience has sometimes been found to moderate the effect of linguistic variables on speech comprehensibility and accentedness. However, since attempts were made to recruit participants, whose backgrounds are the least likely to interfere with results, no moderator analyses were carried out based on demographic variables. Nevertheless, these responses enrich the demographic description of the populations under study and provide necessary information about the sample.

3.6. Data Analysis

The reporting of the results followed the guidelines set out by Larson-Hall and Plonsky (2015), who stress the importance of including information on a) effect sizes and confidence intervals; b) reliability; and c) statistical assumptions, since previous meta-research has found that most manuscripts lack information on one or more of the latter. They argue that the inclusion of such data is important for transparency, replicability, and for future secondary research such as meta-analyses. This is particularly crucial in the case of effect sizes, which can be used by future researchers to conduct power analyses to determine sample size, and by meta-analysts who need effect sizes to pool results from multiple, methodologically diverse studies on the same subject. As for reliability, it is a prerequisite for validity, and without reporting on it, readers cannot be sure whether group differences or other findings are results of research interventions. Finally, statistical assumptions of the inferential tests need to be checked, or else we risk running the wrong tests on the data and drawing inaccurate conclusions.

To determine the effect of CAoA on comprehensibility and accentedness ratings, a procedure similar to the one used by Munro and Derwing (2006) and Alnafisah et al. (2022) was

followed. This entailed first determining the normality of the distributions of the dependent variables for each level of the independent variable. As per the procedures of the above two studies, the native speakers' stimuli were excluded from the inferential statistical tests. SPSS 29 was used to conduct one-way repeated-measures ANOVAs with consonant error type as the within-subjects variable with three levels (early-acquisition error, late-acquisition error, no error). Statistically significant results ($\alpha = .05$) were then followed up with post-hoc tests using Bonferroni correction ($\alpha = .025$). To investigate the relationship between the two dependent variables, statistical assumptions were checked, following which the non-parametric Spearman's correlation test was chosen to account for the non-normality of the distribution.

To determine if a repeated-measures ANOVA could be conducted for the comprehensibility and accentedness scores, the following statistical assumptions were tested: a) outliers; b) normality of distributions for all three levels; and c) sphericity (equality of variances for group differences). The presence of outliers was determined by examining box-and-whisker plots for average comprehensibility ratings at all three levels of the independent variable. These indicated an outlier value in the case of only one participant for comprehensibility ratings for early-acquisition errors. However, the removal of this participant did not change the results of analyses, so it was decided that the data for this participant would be retained. To determine if comprehensibility ratings represented normal distributions, a series of Shapiro-Wilk's tests was conducted. These indicated that comprehensibility scores were normally distributed for no errors ($p = .594$), late-acquisition errors ($p = .849$), and early-acquisition errors ($p = .097$). Finally, to determine if the assumption of sphericity was satisfied, Mauchly's test of sphericity was conducted, which indicated that the assumption has not been violated, $\chi^2(2) = 2.80$, $p = .246$.

Similar to comprehensibility ratings, statistical assumptions were checked to determine the appropriateness of a repeated-measures ANOVA. The presence of outliers was determined through inspecting box-and-whisker plots for average accentedness scores at each of the three levels of the independent variable, which indicated no outliers. This was followed by a series of Shapiro-Wilk's tests to determine the normality of score distributions. These indicated that accentedness ratings were normally distributed for words containing no errors ($p = .392$), for words containing late-acquisition errors ($p = .099$), and for words containing early-acquisition errors ($p = .392$). Finally, Mauchly's test of sphericity was conducted to determine the equality of variances in group differences, which showed that the assumption of sphericity was satisfied, $\chi^2(2)=4.44$, $p = .109$.

In order to determine whether a Pearson's correlation test could be conducted, the following assumptions were checked: a) linearity of the relationship between comprehensibility and accentedness; b) presence of outliers; and c) the normality of the distributions of comprehensibility and accentedness scores. To establish linearity, an observation of the scatterplot displaying the relationship between the two dependent variables indicated that the relationship is linear. Following this, inspection of the box-and-whiskers plots of the two ratings revealed no outliers for either comprehensibility or accentedness scores. Finally, Shapiro-Wilk's tests were run to determine the normality of the distributions and found that neither comprehensibility ratings ($p < .001$), nor accentedness ratings ($p < .001$) exhibited normal distributions. Due to the violation of normality, the decision was made to conduct a non-parametric, two-tailed Spearman's correlation analysis.

Chapter Four: Results

This chapter is divided into three sections. Each of them presents the inferential statistical tests used to answer each of the three research questions. The repeated-measures ANOVAs are followed up with post-hoc comparisons in the case of comprehensibility and accentedness ratings to determine differences at all three levels of the independent variable (No Error, Late-Acquisition Error, Early-Acquisition Error). The correlation between comprehensibility and accentedness is tested using the non-parametric Spearman's correlation.

4.1. Comprehensibility

To determine the effects of CAoA on comprehensibility, a one-way repeated-measures ANOVA was conducted with error type as the within-subjects independent variable with three levels and comprehensibility as the dependent variable. The results of this test indicated that there was a statistically significant effect of error type on comprehensibility scores, $F(2,40) = 135.88$, $p < .001$, $\eta^2 = .87$, which indicates a large effect. Follow-up post-hoc analyses with Bonferroni adjustment were conducted to determine where the differences lay between the three levels. Table 4.1 contains the results of these comparisons. The pairwise comparisons show that comprehensibility ratings were significantly different from each other in the case of all three

Table 4.1

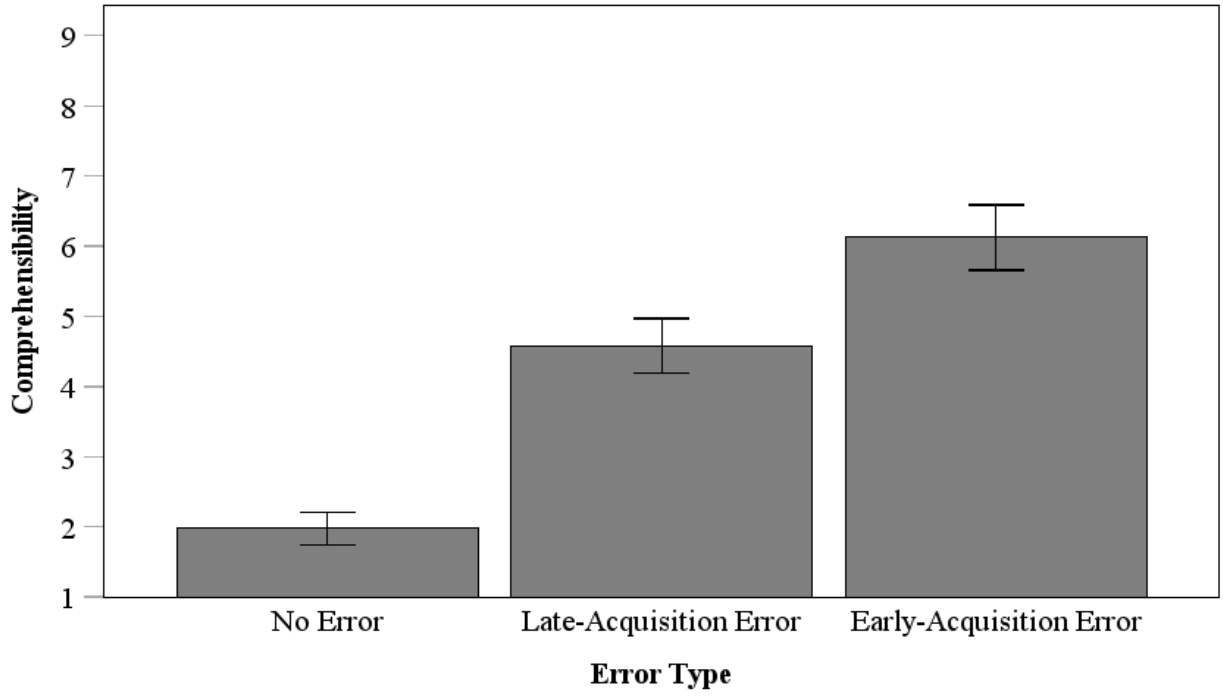
Pairwise Comparisons for Comprehensibility Scores

Baseline	Comparison	Mean Difference [95% *CI*]	*SE*	*p*
No Error	Late	2.60 [2.00, 3.21]	.23	<.001
	Early	4.15 [3.55, 4.75]	.23	<.001
Late	Early	1.54 [.77, 2.32]	.30	<.001

levels of the independent variable. Words were the easiest to understand when they contained no consonant errors (*M* = 1.97, *SD* = 0.64, 95% *CI* [1.74, 2.21]), significantly harder to understand when they contained a late-acquisition error (*M* = 4.58, *SD* = 0.85, 95 % *CI* [4.19, 4.97]), which were in turn significantly easier to understand than early-acquisition errors (*M* = 6.12, *SD* = 1.03, 95% *CI* [5.67, 6.59]). The effects of error type on comprehensibility can also be seen in Figure 4.1. In general, these results confirm the hypothesis that CAoA predicts the relative effects of consonantal errors on Arabic L2 speech comprehensibility. It is important to note that the less than perfect comprehensibility of words containing no errors could stem from certain subphonemic cues, such as stress placement or intonation or even the length of pronouncing certain sounds. For example, the word أُسْوة /ʔuswa/ was pronounced with a slightly longer /s/

Figure 4.1

The Effect of Consonant Error Type on Comprehensibility Scores



Note. Error bars based on 95% *CI*.

sound by the speaker in this condition, and this word received a lower comprehensibility score of 4.95. Such cues, while not considered errors, could nevertheless make it more difficult for listeners to decode the word.

4.2. Accentedness

To determine the effects of CAoA on accentedness, a one-way repeated-measures ANOVA was conducted with error type as the independent variable with three levels, and accentedness as the dependent variable. The results of the analysis indicated that there was a statistically significant effect of error type on accentedness scores, $F(2,40) = 91.44$, $p < .001$, $\eta^2 = .82$, indicating a large effect. This was followed up with post-hoc analyses with Bonferroni adjustment to locate the differences between the three levels. The results of the pairwise comparisons can be seen in Table 4.2. These show that there was a statistically significant difference between each level. Words were considered the least accented when they contained no consonant errors ($M = 3.53$, $SD = 1.57$, 95% CI [2.82, 4.25]), significantly more accented when they contained a late-acquisition error ($M = 6.22$, $SD = 1.53$, 95% CI [5.53, 6.92]), and the most accented when they contained an early-acquisition error ($M = 7.59$, $SD = 1.06$, 95% CI [7.10, 8.07]). These accentedness ratings are graver than comprehensibility scores for all three error

Table 4.2

Pairwise Comparisons for Accentedness Scores

Baseline	Comparison	Mean Difference [95% *CI*]	*SE*	*p*
No Error	Late	2.69 [1.79, 3.59]	.34	<.001
	Early	4.05 [3.18, 4.92]	.33	<.001
Late	Early	1.36 [.77, 1.95]	.23	<.001

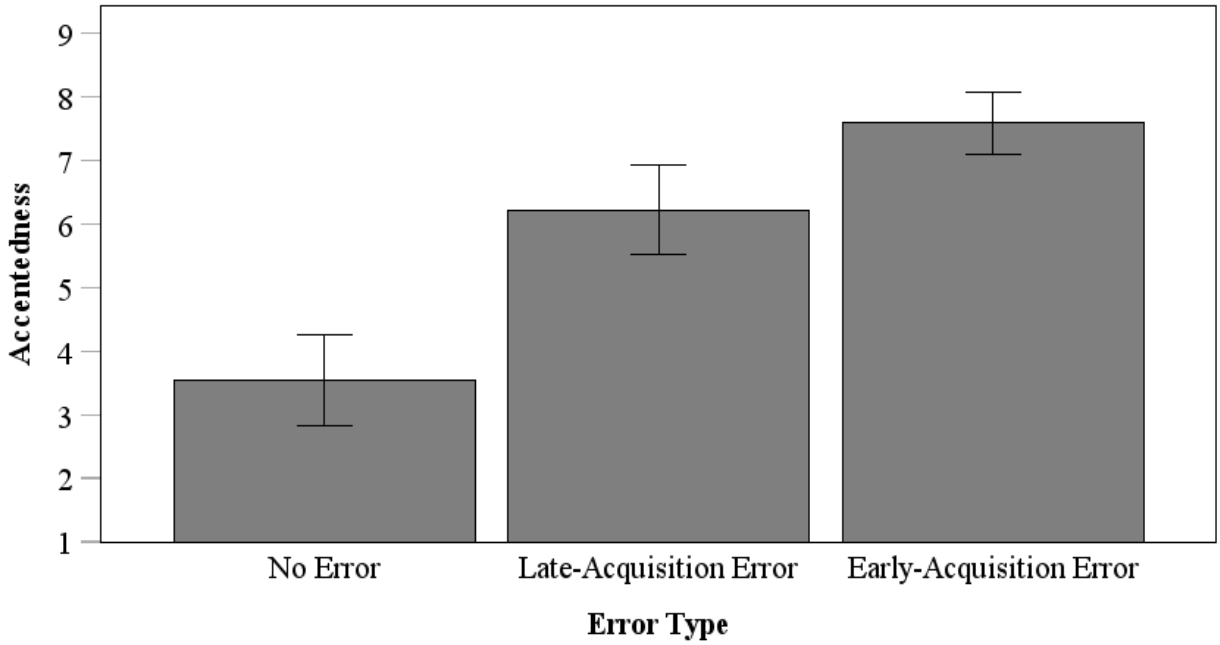
conditions, meaning that words were generally more comprehensible than nativelike, even in cases of no consonant errors. This will be discussed in more detail in the next chapter. However, the mean differences between the conditions are very similar to the ones seen in comprehensibility scores. The effects of error type on accentedness are visually represented in Figure 4.2. In general, the results confirm the hypothesis that CAoA predicts the relative effects of consonantal errors on the perceived degree of foreign-accentedness of L2 Arabic speech.

4.3. The Correlation Between Comprehensibility and Accentedness

Table 4.3 contains the descriptive statistics for averaged overall comprehensibility and accentedness scores. These show that on average, words were rated more comprehensible than nativelike. A Spearman's correlation analysis was conducted to determine the level of

Figure 4.2

The Effect of Consonant Error Type on Accentedness Scores



Note. Error bars based on 95% *CI*.

Table 4.3

Descriptive Statistics for Overall Comprehensibility and Accentedness Scores

Rating	*M* [95% *CI*]	*SD*	*N*
Comprehensibility	4.03 [3.76, 4.30]	3.00	483
Accentedness	5.59 [5.31, 5.86]	3.10	483

association between the two variables, which found a moderate-to-strong correlation between comprehensibility and accentedness scores of individual words by individual listeners, $r_s(481) = .68$, $r_s^2 = .46$, $p < .001$, 95% *CI* [.627, .726]. This means that the variance in comprehensibility ratings explained 46 percent of the variance in accentedness ratings. Generally, words that were harder to understand were also perceived more accented. This result is similar to the one obtained in previous studies testing the correlation between comprehensibility and accentedness.

4.4. Conclusion

The chapter presented the results of the inferential analyses that were used to answer the three research questions. Overall, the results obtained in the study were according to expectations. Comprehensibility scores were most negatively affected by early-acquisition consonant errors, meaning that words containing these errors were the hardest to understand. For example, mispronunciations such as pronouncing عَتْمة /ʕatma/ as أَتْمة [ʔatma] or pronouncing هَيْكَل /hajkal/ as حَيْكَل [ħajkal] made it more difficult for listeners to understand the word than pronouncing صارِم [sˤaːrim] as سارِم [saːrim], pronouncing غَفْلة /ɣafla/ as خَفْلة [xafla] or pronouncing ضَباب /dˤabaːb/ as دَباب [dabaːb]. Similarly, words containing early-acquisition consonant errors were perceived the most foreign-accented. Finally, the correlation between comprehensibility and accentedness was moderate-to-strong, meaning that words that were

harder to understand were generally judged to be more foreign-accented, as well. However, words were generally found to be more comprehensible than nativelike, in keeping with results from previous studies investigating the two ratings. The implications of these results are discussed in the upcoming chapter.

Chapter Five: Discussion

The goal of the present study was to find answers to the following three research questions:

1. Does consonant age of acquisition predict the relative effect of consonantal errors on ratings of speech comprehensibility in L2 Arabic?
2. Does consonant age of acquisition predict the relative effect of consonantal errors on ratings of speech accentedness in L2 Arabic?
3. What is the relationship between L2 Arabic speech comprehensibility and accentedness for read-aloud words?

Accordingly, it built on previous research establishing the predictive power of the FL principle in segmental error gravity in L2 speech. The current study tested a proxy variable (CAoA) that could be used to infer underlying consonant FL distributions in under-researched languages like Arabic. To the best of my knowledge, this is the first investigation of the predictive value of CAoA in segmental error gravity. The results of the current study suggest that CAoA could indeed be used to predict which consonants reduce L2 speech comprehensibility the most. Specifically, consonants that are acquired by L1 ECA-speaking children early seem to affect speech comprehensibility more negatively when mispronounced, compared to consonants that are acquired late. This information could be useful in L2 pronunciation pedagogy, by aiding the development of a hierarchy of mispronunciations that could be relied on to set instructional priorities in the classroom. Similarly, words containing early-acquisition consonant errors were perceived as significantly more foreign-accented than the ones containing late-acquisition errors. This is not surprising in light of evidence suggesting that segmental errors have a sizeable role in listeners' perceptions of foreign-accentedness (Saito, 2021). Furthermore, the correlation

between comprehensibility and accentedness was moderate-to-strong, similar to previous investigations into this relationship. Words were also generally more comprehensible than nativelike.

Besides testing a new variable, the current investigation also incorporated some methodological innovations. Firstly, it is the first study I know of that controlled for the position of the mispronounced segments: consonant errors only occurred in word-initial position. Secondly, it controlled for word length through the selection of disyllabic items. Thirdly, grammatical category was also controlled for to a large extent through the sole inclusion of content words (mostly nouns and some adjectives). Lastly, the potential confounding effect of real words was mitigated by excluding all mispronunciations that resulted in recognizable lexical items in Arabic. Overall, read-aloud words seem to be a viable, more controlled way of investigating segmental error gravity. Of course, despite the promising results, the study also has some limitations, such as the incomplete representation of consonant errors in the data and the uneven distribution of the speakers across the error conditions. In the following sections, the findings are compared and contrasted with previous studies exploring segmental error gravity in other languages. Theoretical and practical implications are discussed in their dedicated sections. Finally, limitations are addressed and directions for further research are offered in the domain.

5.1. The Effects of Consonant Error Type

While the variable CAoA has not been investigated before in L2 pronunciation studies, the results of this study can nevertheless be compared to those of previous studies on the effects of the FL of phoneme substitutions. This can be done with the assumption that early-acquisition consonants represent high FL and late-acquisition ones represent low FL in the current study. Importantly, while previous studies have looked at the FL value of the phoneme pair involved in

the substitution (i.e., the original sound and the sound that was substituted in), the current study assumed FL values for individual consonants, these being the original consonants in the word. This was probably not an issue, since the substitutions in the current study involved consonants belonging to the same category, meaning that early-acquisition consonants were replaced with other early-acquisition ones by speakers, and late-acquisition consonants were likewise replaced with late-acquisition ones. This conceptualization of FL goes beyond minimal pairs and includes non-word competitors, in alignment with Sewell's (2021) recommendation. Such an approach allows us to study the mispronunciation of individual sounds and not rely on a list of most commonly substituted phoneme pairs. This approach could also possibly give way to the study of errors beyond substitutions, such as distortions and deletions of individual phonemes, where the resulting pronunciation does not involve a target language phoneme.

The results of the present investigation replicate the findings of Munro and Derwing (2006) and Alnafisah et al. (2022) comparing the effects of low and high FL substitutions on comprehensibility and accentedness. In both of these studies, error type had a large effect on both comprehensibility and accentedness ratings: high FL errors resulted in more negative comprehensibility and stronger accentedness ratings. Out of the two studies, the effect sizes of the current investigation align with Munro and Derwing (2006), which are much larger than the ones reported by Alnafisah et al. (2022). This is probably due to the methodological choice of excluding sentences containing no errors from the ANOVAs in the case of Alnafisah et al., since the jump from no errors to one low FL error shows the steepest change in comprehensibility and accentedness scores. This change of the latter type in comprehensibility ratings seems even steeper in the current study compared to Munro and Derwing (2006), which could probably be

attributed to the decontextualized nature of the mispronounced words. However, the general pattern of the results seems to follow the ones attested in these two studies.

The results of the present study are also similar to the ones employing a continuous operationalization of FL, in the form of the ratio of the mispronounced segments compared to the total number of segments in a sentence. These studies only looked at the correlation between comprehensibility scores and the ratio of the different types of segmental errors. Suzukida and Saito (2021) similarly found that high FL consonant error ratio was moderately strongly correlated with lower comprehensibility in two studies examining extemporaneous L2 English speech. However, in their studies, low FL consonant error ratio was not significantly associated with comprehensibility scores. This could be due to the methodological difference in the operationalization of error type. Alternatively, it could have arisen from the nature of the speaking task, since this has been the only study that used extemporaneous speech instead of a read-aloud task. On the other hand, Bao et al.'s (2022) investigation of FL effects on the comprehensibility of L2 Mandarin read-aloud sentences found a moderate-to-strong correlation between high FL error ratio and a relatively weak, but significant correlation between low FL error ratio and comprehensibility. While the results of the latter two studies might not be directly comparable to the current ones, the tendencies uncovered in the present investigation point to the same direction.

5.2. The Relationship Between Comprehensibility and Accentedness

The relationship between comprehensibility and accentedness has mostly been compared in studies that also included measures of intelligibility. Many of those studies utilized mixed-effects and hierarchical regression models and cannot be compared directly to the current study. However, the correlation between comprehensibility and accentedness can be compared to both

the one found for extemporaneous L2 Arabic speech and the one found for individual words. To the best of my knowledge, the current study is the first one that measured this correlation for read-aloud words, since the one study using read-aloud words only measured accentedness, intelligibility, and reaction time, but excluded comprehensibility (Caspers & Horłoza, 2012). In addition, the other study using individual words employed picture-elicited production of the lexical items (Uchihara, 2022). For this reason, while previous studies on FL effects did not include tests of correlation, it was important to establish this relationship.

In the case of L2 Arabic speech, the current results found a similar correlation to the one reported by Ali (2023). In his study, the correlation was strong (r = .75) between comprehensibility and accentedness ratings of the speakers, compared to the moderate-to-strong (r = .68) relationship found in the current investigation. However, the 95% confidence interval reported by Ali is relatively wide, ranging from .49 to 1.00, indicating a large margin of error that includes all values from moderate to perfect correlation. This could be due to the way the correlation was computed, which was the correlation between comprehensibility and accentedness ratings averaged across listeners for each speaker. In contrast, the current study calculated correlation between each individual comprehensibility and accentedness rating, resulting in a large number of observations, which reduces error.

The relationship observed between comprehensibility and accentedness is also very similar to the one uncovered by Uchihara (2022) for picture-elicited individual words in L2 English. In his study, the correlation between comprehensibility and accentedness ratings was .67, which is within the confidence interval of the result found in the current study. Furthermore, his calculation was based on the individual ratings from each listener given to each word, which is the methodology that the current study employed. Thus, the current finding further supports

the notion that comprehensibility and accentedness are related, but independent dimensions of L2 speech even at the level of individual words, and in a language other than English.

In general, the graver accentedness ratings compared to the comprehensibility scores make sense from the standpoint of the differences in the constructs. While comprehensibility solely measures the effort needed to understand speech, accentedness requires the identification of the degree of foreignness of speech. As such, even speakers that are relatively easy to understand could be perceived to have a large degree of foreign accent, due to certain segmental cues like vowel quality (listeners seem to pay attention to segments when judging foreign accent; Saito, 2021). This could be the true even if these cues do not constitute mispronunciations, which is evident in the case of the word حافِز /ħa:fiz/ that received a comprehensibility score of 3.43, but an accentedness rating of 7.76. The speaker pronounced this word with a closer /i/ vowel, which is not a segmental error, but sounds markedly non-native.

5.3. Theoretical and Methodological Implications

The results obtained in this investigation seem to lend preliminary support to the assumption that CAoA is correlated with FL in the case of Arabic. Of course, since CAoA was not operationalized as a continuous variable, no fine-grained comparisons can be made. In addition, while L1 phonological acquisition studies tend to categorize phonemes into early, middle, and late stages of acquisition, the current investigation used an early and late distinction, in line with L2 pronunciation studies comparing errors having high FL and low FL. This was necessary for the sake of comparability between the present results and previous studies using FL. Future L1 acquisition-oriented studies would benefit from operationalizing both FL and CAoA as continuous variables to uncover detailed ambient language effects on children acquiring Arabic as a native language.

When it comes to using CAoA as a predictor of segmental error gravity in other under-researched languages, the present study offers a promising framework for doing that. Firstly, researchers should look at the phonological structure of the language in question to determine whether consonants have a relatively high contrastive role compared to vowels and suprasegmental features. For example, CAoA would not be a suitable candidate for predicting segmental error gravity in languages possessing vowel-heavy phoneme inventories or relying on tones to contrast lexical meaning. Secondly, psycholinguistic evidence should be examined that compares the role that consonants play in both children's and adults' speech processing. The more domains a consonant bias emerges in and the earlier it emerges, the higher the likelihood that CAoA is strongly associated with FL in that language.

The present study also employed a new methodology in the form of using read-aloud words to investigate segmental error gravity. While originally utilized by Caspers and Horłoza (2012) to compare the gravity of segmental and suprasegmental errors in L2 Dutch, the method seems to be appropriate for a controlled comparison of different types of segmental errors. Read-aloud words afford researchers a high level of experimental control by ensuring that only the targeted errors take place in the stimuli and ensuring that stimuli length, error position and extra-phonemic factors do not interfere with the comparison. Of course, it could happen that once linguistic context or additional factors are added, the effects of certain types of errors are canceled out (Thir, 2020). For this reason, read-aloud words are but one level of analysis that should be used in tandem with other, more ecologically valid elicitation methods to triangulate pronunciation error gravity.

5.4. Pedagogical Implications

The present study aimed to find an organizing principle that could be used to determine segmental error gravity in AFL. This was done in place of available data on the FL distribution of Arabic phonemes, by making an indirect inference about likely FL distributions on the basis of available CAoA data. In L2 pronunciation pedagogy, setting priorities is important because of the exceptional difficulty that pronunciation poses to L2 learners. Instructors simply do not have time to focus on all mispronunciations in class, and doing so is a futile effort. Insisting on nativeness standards in L2 phonology also does not respect student's autonomy to retain non-native identities when speaking the target language, since accents are powerful social markers. At the same time, not explicitly addressing pronunciation would be a mistake since some mispronunciations affect speech comprehensibility and intelligibility very negatively and thus cause breakdowns in communication and avoidance to interact on the part of interlocutors. Leaving these errors unaddressed does a disservice to students who might not even be aware of the sources of the breakdowns they experience in communication. Errors also do not necessarily resolve on their own without instructional focus just by mere increases in proficiency (Derwing, 2017). This is evidenced in the fact that the speakers in the current study were all at the Advanced and Superior levels of proficiency according to their self-reports.

The results of the current study thus provide the first empirical evidence on what the most serious pronunciation errors in Arabic might be. In general, the current study found that early-acquisition consonants are potentially more important for communication in Arabic than late-acquisition ones. Early-acquisition consonants were represented by ح and ه, while the late-acquisition category included خ, غ, ص, ض. The fact that mispronouncing the latter group seemed to affect listener understanding less also makes sense from a sociolinguistic point of view. The latter group of consonants is subject to more inter-dialectal variation: e.g., ض is entirely absent

from Peninsular dialects that replace it with ظ. It is also replaced with د in certain roots in ECA (e.g., ضحك). The same way, ص is substituted with س in some words in ECA, such as سِدر.

It is important to point out that a segmental error gravity hierarchy applies to phonological production and not to perception. The relationship between segmental perception and production remains unclear, with studies having found conflicting results (Nagle & Baese-Berk, 2022). In L2 Arabic, perception accuracy seems to be higher than production accuracy (Al Tubuly, 2018). In general, there is no reason to assume that perception training will automatically carry over to production. It could also be assumed that the priorities for perception training will be different since it belongs in the domain of listening instruction. A case could be made for the relatively equal importance of segments in perception because AFL learners mostly listen to native speech and need to adequately discriminate between sounds for listening comprehension. However, in speech production, the intelligibility principle represents a framework of selective accuracy where accuracy is only pursued for features that hurt speech intelligibility and comprehensibility the most. This is both in contrast with the nativeness principle corresponding to accuracy and the natural approach that favors fluency over accuracy.

5.5. Limitations and Directions for Future Research

Naturally, the results of this study should be taken as preliminary. While the results support the hypothesis that early-acquisition consonant errors affect speech comprehensibility more negatively, the errors under investigation did not include all consonants within the two categories. Early-acquisition errors involved the substitution of ه /h/ and ع /ʕ/ but did not include ح /ħ/ and ء /ʔ/. While the latter two mispronunciations did occur in the recorded word list, they had to be excluded later because it was discovered that they either resulted in a real word, or contained a vowel that could be considered non-native by Egyptian listeners, which would have

confounded the results. Late-acquisition errors involved the substitution of خ /x/, غ /ɣ/, ص /sˤ/ and ض /dˤ/, but did not include ط /tˤ/, ظ /ðˤ/, and ق /q/. This was due to the nature of the word list. Since only medium-frequency nouns and adjectives were selected in a way to only include one of the 11 potentially difficult consonants in each word, this restricted the candidates for word selection. Future studies could resolve this by expanding the scope of word inclusion, for example by including high-frequency items, so that each consonant can be represented enough times for mispronunciations to occur.

In an ideal situation, speakers would be equally represented across error conditions, which was not the case in the current study. The early-acquisition errors were made by three speakers, while the late-acquisition errors were committed by four speakers. While the no error condition featured all ten speakers, overall, only seven speakers contributed with mispronounced tokens in the analyses. Although other errors were controlled for, other subphonemic cues could still affect ratings. Ideally, each error condition would include tokens from each speaker, to maximize the comparability of the conditions. For such a comparison, experimental manipulations seem especially appropriate, such as manually replacing mispronounced segments with correct ones, or the other way round, manually including mispronunciations in originally correctly produced words. This method has been successfully utilized to compare the relative contribution of segments and suprasegmentals, by transferring segments and intonation between native and non-native speakers' recordings (Sereno et al., 2016; Yang et al., 2021). Such an approach essentially amounts to direct experimental manipulation and would further increase the validity of the findings.

It is important to stress that the current investigation does not make claims about the importance of consonantal mispronunciations relative to other aspects of speech, but merely

compares consonant errors to each other. However, based on the psycholinguistic evidence reviewed earlier that establishes the predominance of consonants in Arabic speech perception, it is likely that consonants are more important than vowels for speech intelligibility and comprehensibility. This prediction is further reinforced by the significant regional and social variation in the realization of short vowels. For example, the word شُبّاك /ʃubːaːk/ is realized as شِبّاك [ʃibːaːk] in Egyptian Arabic. However, there might be a limit to this tolerance for vowel variation. For instance, many North African dialects exhibit the reduction of short vowels to a schwa [ə], and these dialects tend to be considered unintelligible by speakers of Eastern dialects. Potentially, intelligibility studies could employ native speech evaluated by native listeners to enrich our knowledge base.

Furthermore, the study did not investigate other aspects of speech production, such as suprasegmental features, fluency (speaking rate) and lexicogrammatical factors. While research has uncovered the effect of these on L2 English comprehensibility, they remain to be explored in the case of Arabic. The suprasegmental features of vowel length and gemination could certainly be investigated in future studies. As for fluency, there is no reason to believe that previous results about an optimal speaking rate do not apply in the case of Arabic: speech that is too slow or too fast is probably harder to understand. Lexical and grammatical accuracy and complexity could also potentially have an effect and are best explored through extemporaneous speaking tasks, such as the Oral Proficiency Interview. In Arabic, the latter factors could be complicated by diglossia, and it would be useful to uncover the effect of, for example, diglossic mixing in vocabulary and grammar on speech comprehensibility. Such mixing tends to be common among learners of Arabic, and results in lexicogrammatical choices that might be very different from the way native speakers mix colloquial and standard.

While the current investigation has focused on the what and the why of L2 Arabic pronunciation, it does not explicitly address the how. Pedagogical practices need to undergo empirical scrutiny in the form of longitudinal classroom research. Crucially, any pronunciation gains can only be verified through listener measures, i.e., improvements in intelligibility and comprehensibility over time, since these dimensions are what determine the likely success of spoken interactions (Derwing & Munro, 2022). Again, this means a framework of selective accuracy, which is somewhat different from the framework proposed by Rifaat (2017) for AFL, which is informed by Krashen's natural approach and proposes fluency-focused instruction in the lower levels of proficiency. However, since pronunciation errors might not resolve on their own over time, it seems more beneficial to target the gravest errors in the classroom from the very beginning. Early intervention would also help prevent the fossilization of these errors.

Finally, listener ratings of comprehensibility and accentedness should be collected from multiple dialects to further explore the segmental error hierarchy in Arabic. The first obvious candidate would be the Syrian dialect, since the CAoA data available are identical to the data in Egyptian Arabic. However, attention should be paid to only include word forms that are shared between MSA and the colloquial dialect under study. Speakers of dialects seem to be sensitive to voweling that does not agree with their mental representation of the vowels of words, even if said voweling seems to correspond to the MSA dictionary form of the word. In a similar vein, research involving listeners from specific dialect backgrounds could also investigate the effect of errors in dialect-specific phonological features, such as the substitution of the emphatic [rˤ] (e.g., in the word [rˤɑːgil]) with the non-emphatic [r] (e.g., in the word [ræːgiʔ]) in Egyptian Arabic, and vice versa. Evidence about the features affecting the intelligibility and comprehensibility of

L2 colloquial speech would be immensely useful in a time when colloquial Arabic is being increasingly common in university classrooms and curricula.

5.6. Conclusion

The present investigation aimed to find an organizing principle for Arabic segmental error gravity. Specifically, it tested whether the age of acquisition of consonantal errors predicts reduced comprehensibility and higher levels of foreign-accentedness. According to expectations, the results show that mispronouncing early-acquisition consonants does have a more detrimental effect on speech comprehensibility and it does increase the perceived degree of foreign-accentedness. In addition, the study showed that comprehensibility and accentedness are related, but independent dimensions of L2 Arabic speech even for read-aloud words. While words that were harder to understand were also generally found to be more accented, there were exceptions for some words rated by some listeners. The theoretical significance of the study is that it showed that alternative ways of predicting segmental error gravity are possible in the case of under-researched languages for which only developmental data are available. The main pedagogical takeaway should be that pronunciation does not necessarily improve on its own without instructional focus, evidenced by the mispronunciations that occur even at the Advanced and Superior levels of proficiency. Furthermore, this instructional focus should be dedicated to mispronunciations that are more detrimental to speech intelligibility and comprehensibility. Further research should aim to triangulate the promising results of this study, through including a wider range of errors from a wider range of speakers and testing them on listeners from multiple dialect backgrounds.

References

Ahmed, A., Ali, N., Alzubaidi, M., Zaghouani, W., Abd-alrazaq, A. A., & Househ, M. (2022). Freely available Arabic corpora: A scoping review. *Computer Methods and Programs in Biomedicine Update*, *2*, 100049. https://doi.org/10.1016/j.cmpbup.2022.100049

Al Tubuly, S. (2018). The perception and production of Arabic consonants: A cross-linguistic study. In M. T. Alhawary (Ed.), *Routledge handbook of Arabic second language acquisition* (pp. 70–92). Taylor & Francis Group. https://doi.org/10.4324/9781315674261

Aldholmi, Y., & Pycha, A. (2023). Segmental contributions to word recognition in Arabic sentences. *Poznan Studies in Contemporary Linguistics*, *59*(2), 257–287. https://doi.org/10.1515/psicl-2022-2010

Ali, M. M. E. (2023). The foreign-accentedness, comprehensibility, and intelligibility of L2 Arabic speech. *Language Teaching Research*, Advance online publication. https://doi.org/10.1177/13621688231158787

Alnafisah, M., Goodale, E., Rehman, I., Levis, J. M., & Kochem, T. (2022). The impact of functional load and cumulative errors on listeners' judgments of comprehensibility and accentedness. *System*, *110*, Article 102906. https://doi.org/10.1016/j.system.2022.102906

Amayreh, M. M. (2003). Completion of the consonant inventory of Arabic. *Journal of Speech, Language, and Hearing Research*, *46*(3), 517–529. https://doi.org/10.1044/1092-4388(2003/042)

Amayreh, M. M., & Dyson, A. T. (1998). The acquisition of Arabic consonants. *Journal of Speech, Language & Hearing Research*, *41*(3), 642. https://doi.org/10.1044/jslhr.4103.642

Amayreh, M. M., & Dyson, A. T. (2000). Phonetic inventories of young Arabic-speaking children. *Clinical Linguistics & Phonetics*, *14*(3), 193–215. https://doi.org/10.1080/026992000298823

Appel, R., Trofimovich, P., Saito, K., Isaacs, T., & Webb, S. (2019). Lexical aspects of comprehensibility and nativeness from the perspective of native-speaking English raters. *ITL - International Journal of Applied Linguistics*, *170*(1), 24–52. https://doi.org/10.1075/itl.17026.app

Archibald, J. (2017). Transfer, contrastive analysis and interlanguage phonology. In O. Kang, R. I. Thomson, & J. M. Murphy (Eds.), *The Routledge handbook of contemporary english pronunciation* (pp. 8–24). Routledge. https://doi.org/10.4324/9781315145006-2

Archibald, J. (2021). Ease and difficulty in L2 phonology: A mini-review. *Frontiers in Communication*, *6*. https://www.frontiersin.org/articles/10.3389/fcomm.2021.626529

Badawi, E.-S. M., & Hinds, M. (1986). *A dictionary of Egyptian Arabic: Arabic-English*. Librairie du Liban.

Baker, A. A. (2017). Pronunciation teaching in the pre-CLT era. In O. Kang, R. I. Thomson, & J. M. Murphy (Eds.), *The Routledge handbook of contemporary English pronunciation* (pp. 248–266). Routledge. https://doi.org/10.4324/9781315145006-16

Bao, R., Peng, L., & Zhang, J. (2022). The contributions of initials and finals to L2 Chinese comprehensibility based on functional load principle. *2022 International Conference on Asian Language Processing (IALP)*, 358–362. https://doi.org/10.1109/IALP57159.2022.9961277

Baugh, J. (2017). Linguistic profiling and discrimination. In O. Garcia, N. Flores, & M. Spotti (Eds.), *The Oxford handbook of language and society* (pp. 349–367). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190212896.001.0001

Behrens, H. (2009). Usage-based and emergentist approaches to language acquisition. *Linguistics*, *47*(2), 383–411. https://doi.org/10.1515/LING.2009.014

Bent, T., Bradlow, A. R., & Smith, B. L. (2007). Segmental errors in different word positions and their effects on intelligibility of non-native speech. In O.-S. Bohn & M. J. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege*. John Benjamins Publishing.

Bouchon, C., Hochmann, J.-R., & Toro, J. M. (2022). Spanish-learning infants switch from a vowel to a consonant bias during the first year of life. *Journal of Experimental Child Psychology*, *221*, 105444. https://doi.org/10.1016/j.jecp.2022.105444

Brown, A. (1988). Functional load and the teaching of pronunciation. *TESOL Quarterly*, *22*(4), 593–606. https://doi.org/10.2307/3587258

Buckwalter, T., & Parkinson, D. (2011). *A frequency dictionary of Arabic: Core vocabulary for learners*. Routledge.

Caldwell-Harris, C. L., & MacWhinney, B. (2023). Age effects in second language acquisition: Expanding the emergentist account. *Brain and Language*, *241*, 105269. https://doi.org/10.1016/j.bandl.2023.105269

Caspers, J., & Horłoza, K. (2012). Intelligibility of non-natively produced Dutch words: Interaction between segmental and suprasegmental errors. *Phonetica*, *69*(1–2), 94–107. https://doi.org/10.1159/000342622

Catford, J. C. (1987). Phonetics and the teaching of pronunciation: A systemic description of English phonology. In J. Morley (Ed.), *Current perspectives on pronunciation: Practices anchored in theory* (pp. 87–100). Teachers of English to Speakers of Other Languages.

Chen, H., Lee, D. T., Luo, Z., Lai, R. Y., Cheung, H., & Nazzi, T. (2021). Variation in phonological bias: Bias for vowels, rather than consonants or tones in lexical processing by Cantonese-learning toddlers. *Cognition*, *213*, 104486. https://doi.org/10.1016/j.cognition.2020.104486

Crowther, D., Holden, D., & Urada, K. (2022). Second language speech comprehensibility. *Language Teaching*, *55*(4), 470–489. https://doi.org/10.1017/S0261444821000537

Crowther, D., & Urada, K. (2022). *Face-to-face versus remote L2 speech elicitation: Listeners' perceptions of sound clarity*. https://www.hawaii.edu/sls/wp-content/uploads/Crowther-Urada-2022.pdf

Cychosz, M. (2017). Functional load and frequency predict consonant emergence across five languages. *UC Berkeley PhonLab Annual Report*, *13*(1). https://doi.org/10.5070/P7131040758

Derwing, T. M. (2017). The efficacy of pronunciation instruction. In O. Kang, R. I. Thomson, & J. M. Murphy (Eds.), *The Routledge handbook of contemporary English pronunciation*. Routledge. https://doi.org/10.4324/9781315145006

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. John Benjamins Publishing Company.

Derwing, T. M., & Munro, M. J. (2022). Pronunciation learning and teaching. In T. M. Derwing, M. J. Munro, & R. I. Thomson (Eds.), *The Routledge handbook of second language*

acquisition and speaking (pp. 147–159). Routledge. https://doi.org/10.4324/9781003022497

Derwing, T. M., Rossiter, M. J., & Munro, M. J. (2002). Teaching native speakers to listen to foreign-accented speech. *Journal of Multilingual and Multicultural Development*, *23*(4), 245–259. https://doi.org/10.1080/01434630208666468

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. https://doi.org/10.1111/j.1467-9922.2004.00282.x

Diessel, H. (2017). Usage-based linguistics. In *Oxford research encyclopedia of linguistics*. Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.363

Dragojevic, M., & Goatley-Soan, S. (2022). Americans' attitudes toward foreign accents: Evaluative hierarchies and underlying processes. *Journal of Multilingual and Multicultural Development*, *43*(2), 167–181. https://doi.org/10.1080/01434632.2020.1735402

Eckman, F. R. (2008). Typological markedness and second language phonology. In J. G. Hansen Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (Vol. 36, pp. 95–115). John Benjamins Publishing Company. https://doi.org/10.1075/sibil.36.06eck

Elrefaie, D. A., Hegazi, M. A. E.-F., El-Mahallawi, M. M., & Khodeir, M. S. (2021). Descriptive analysis of the development of the Arabic speech sounds among typically developing colloquial Egyptian Arabic-speaking children. *The Egyptian Journal of Otolaryngology*, *37*, Article 38. https://doi.org/10.1186/s43163-021-00094-w

Flege, J. E. (1984). The detection of French accent by American listeners. *The Journal of the Acoustical Society of America*, *76*(3), 692–707. https://doi.org/10.1121/1.391256

Froud, K., & Khamis-Dakwar, R. (2021). The study of Arabic language acquisition: A critical review. In K. Ryding & D. Wilmsen (Eds.), *The Cambridge handbook of Arabic linguistics* (pp. 48–82). Cambridge University Press. https://doi.org/10.1017/9781108277327.004

Gass, S. M., & Varonis, E. M. (1985). Variation in native speaker speech modification to non-native speakers. *Studies in Second Language Acquisition*, *7*(1), 37–57. https://www.jstor.org/stable/44488545

Gluszek, A., & Dovidio, J. F. (2010). The way they speak: A social psychological perspective on the stigma of nonnative accents in communication. *Personality and Social Psychology Review*, *14*(2), 214–237. https://doi.org/10.1177/1088868309359288

Gwasmeh, M. (2021). *Language attitudes, accentedness and comprehensibility: A sociolinguistic study of Arabic and Arabic-accented English.* [Doctoral dissertation, University of Canterbury]. https://ir.canterbury.ac.nz/handle/10092/102380

Gwilliams, L., & Marantz, A. (2015). Non-linear processing of a linear speech stream: The influence of morphological structure on the recognition of spoken Arabic words. *Brain and Language*, *147*, 1–13. https://doi.org/10.1016/j.bandl.2015.04.006

Hellmuth, S. (2014). Towards a research-led approach to the teaching of Arabic pronunciation. *Proceedings of the International Symposium on the Acquisition of Second Language Speech Concordia Working Papers in Applied Linguistics*, *5*, 295–309. http://doe.concordia.ca/copal/documents/21_Hellmuth_Vol5.pdf

Holliday, A. (2006). Native-speakerism. *ELT Journal*, *60*(4), 385–387. https://doi.org/10.1093/elt/ccl030

Ingram, D. (2008). Cross-linguistic phonological acquisition. In M. J. Ball, M. R. Perkins, N. Mller, & S. Howard (Eds.), *The handbook of clinical linguistics* (pp. 626–640). Blackwell Publishing Ltd. https://doi.org/10.1002/9781444301007.ch38

Isaacs, T. (2009). Integrating form and meaning in L2 pronunciation instruction. *TESL Canada Journal*, *27*(1), 1–12. http://dx.doi.org/10.18806/tesl.v27i1.1034

Jułkowska, I. A., & Cebrian, J. (2015). Effects of listener factors and stimulus properties on the intelligibility, comprehensibility and accentedness of L2 speech. *Journal of Second Language Pronunciation*, *1*(2), 211–237. https://doi.org/10.1075/jslp.1.2.04jul

Kastner, I., & Adriaans, F. (2018). Linguistic constraints on statistical word segmentation: The role of consonants in Arabic and English. *Cognitive Science*, *42*(S2), 494–518. https://doi.org/10.1111/cogs.12521

King, R. D. (1967). Functional load and sound change. *Language*, *43*(4), 831–852. https://doi.org/10.2307/411969

Lador-Weizman, Y., & Deutsch, A. (2022). The contribution of consonants and vowels to auditory word recognition is shaped by language-specific properties: Evidence from Hebrew. *Journal of Experimental Psychology: Human Perception and Performance*, *48*(5), 401–426. https://doi.org/10.1037/xhp0001000.supp

Larson-Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning*, *65*(S1), 127–159. https://doi.org/10.1111/lang.12115

Lev-Ari, S., Ho, E., & Keysar, B. (2018). The unforeseen consequences of interacting with non-native speakers. *Topics in Cognitive Science*, *10*(4), 835–849. https://doi.org/10.1111/tops.12325

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, *39*(3), 369–377. https://doi.org/10.2307/3588485

Levis, J. M. (2020). Revisiting the intelligibility and nativeness principles. *Journal of Second Language Pronunciation*, *6*(3), 310–328. https://doi.org/10.1075/jslp.20050.lev

Levis, J. M., & Sonsaat, S. (2017). Pronunciation teaching in the early CLT era. In O. Kang, R. I. Thomson, & J. M. Murphy (Eds.), *The Routledge handbook of contemporary English pronunciation* (pp. 267–283). Routledge. https://doi.org/10.4324/9781315145006-17

Ma, W., Zhou, P., Singh, L., & Gao, L. (2017). Spoken word recognition in young tone language learners: Age-dependent effects of segmental and suprasegmental variation. *Cognition*, *159*, 139–155. https://doi.org/10.1016/j.cognition.2016.11.011

Major, R. C. (2007). Identifying a foreign accent in an unfamiliar language. *Studies in Second Language Acquisition*, *29*(4), 539–556. https://doi.org/10.1017/S0272263107070428

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, *27*(7–8), 953–978. https://doi.org/10.1080/01690965.2012.705006

McLeod, S., & Crowe, K. (2018). Children's consonant acquisition in 27 languages: A cross-linguistic review. *American Journal of Speech-Language Pathology*, *27*(4), 1546–1571. https://doi.org/10.1044/2018_AJSLP-17-0100

Moyer, A. (2013). *Foreign accent: The phenomenon of non-native speech*. Cambridge University Press.

Munro, M. J., & Derwing, T. M. (1995a). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45*(1), 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Munro, M. J., & Derwing, T. M. (1995b). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech*, *38*(3), 289–306. https://doi.org/10.1177/002383099503800305

Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech: The role of speaking rate. *Studies in Second Language Acquisition*, *23*(4), 451–468. https://doi.org/10.1017/S0272263101004016

Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System*, *34*(4), Article 102906. https://doi.org/10.1016/j.system.2006.09.004

Munro, M. J., & Derwing, T. M. (2015). A prospectus for pronunciation research in the 21st century: A point of view. *Journal of Second Language Pronunciation*, *1*(1), 11–42. https://doi.org/10.1075/jslp.1.1.01mun

Munro, M. J., & Derwing, T. M. (2020). Foreign accent, comprehensibility and intelligibility, redux. *Journal of Second Language Pronunciation*, *6*(3), 283–309. https://doi.org/10.1075/jslp.20038.mun

Murphy, J. M., & Baker, A. A. (2015). History of ESL pronunciation teaching. In M. Reed & J. M. Levis (Eds.), *The handbook of English pronunciation* (pp. 36–65). Wiley. https://doi.org/10.1002/9781118346952.ch3

Na, Y. (2021). *The impact of segmental accuracy on intelligibility* [Master's thesis, University of Hawai'i]. https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/a8b77f14-7b84-4ca7-a2e3-0fe3bdd0a30b/content

Nagle, C. L., & Baese-Berk, M. M. (2022). Advancing the state of the art in L2 speech perception-production research: Revisiting theoretical assumptions and methodological

practices. *Studies in Second Language Acquisition*, *44*(2), 580–605. https://doi.org/10.1017/S0272263121000371

Nagle, C. L., & Huensch, A. (2020). Expanding the scope of L2 intelligibility research: Intelligibility, comprehensibility, and accentedness in L2 Spanish. *Journal of Second Language Pronunciation*, *6*(3), 329–351. https://doi.org/10.1075/jslp.20009.nag

Nazzi, T., & Cutler, A. (2019). How consonants and vowels shape spoken-language recognition. *Annual Review of Linguistics*, *5*(1), 25–47. https://doi.org/10.1146/annurev-linguistics-011718-011919

Neal, R. (2022). How do L1 Chinese raters process the L2 Chinese speech signal at the sentence level with respect to accentedness, comprehensibility and intelligibility? *Chinese as a Second Language Research*, *11*(2), 233–288. https://doi.org/10.1515/caslar-2022-2003

Owaida, H. (2015). *Speech sound acquisition and phonological error patterns in child speakers of Syrian Arabic: A normative study* [Unpublished PhD dissertation, City University London]. https://openaccess.city.ac.uk/id/eprint/15182/

Purnell, T., Idsardi, W., & Baugh, J. (1999). Perceptual and phonetic experiments on American English dialect identification. *Journal of Language and Social Psychology*, *18*(1), 10–30. https://doi.org/10.1177/0261927X99018001002

Ratnage, P., Nazzi, T., & Floccia, C. (2023). Vowels and consonants matter equally to British English-learning 11-month-olds' familiar word form recognition. *Journal of Child Language*, 1–24. https://doi.org/10.1017/S0305000923000223

Rifaat, K. (2017). A strategy for teaching Arabic pronunciation. In K. M. Wahba, L. England, & Z. A. Taha (Eds.), *Handbook for Arabic language teaching professionals in the 21st century* (Vol. 2, pp. 329–343). Routledge. https://doi.org/10.4324/9781315676111

Saito, K. (2021). What characterizes comprehensible and native-like pronunciation among English-as-a-second-language speakers? Meta-analyses of phonological, rater, and instructional factors. *TESOL Quarterly*, *55*(3), 866–900. https://doi.org/10.1002/tesq.3027

Sereno, J., Lammers, L., & Jongman, A. (2016). The relative contribution of segments and intonation to the perception of foreign-accented speech. *Applied Psycholinguistics*, *37*(2), 303–322. https://doi.org/10.1017/S0142716414000575

Severen, L. V., Gillis, J. J. M., Molemans, I., Berg, R. V. D., Maeyer, S. D., & Gillis, S. (2013). The relation between order of acquisition, segmental frequency and function: The case of word-initial consonants in Dutch. *Journal of Child Language*, *40*(4), 703–740. https://doi.org/10.1017/S0305000912000219

Sewell, A. (2017). Functional load revisited: Reinterpreting the findings of 'lingua franca' intelligibility studies. *Journal of Second Language Pronunciation*, *3*(1), 57–79. https://doi.org/10.1075/jslp.3.1.03sew

Sewell, A. (2021). Functional load and the teaching-learning relationship in L2 pronunciation. *Frontiers in Communication*, *6*, Article 627378. https://www.frontiersin.org/articles/10.3389/fcomm.2021.627378

Stokes, S. F., & Surendran, D. (2005). Articulatory complexity, ambient frequency, and functional load as predictors of consonant development in children. *Journal of Speech, Language, and Hearing Research*, *48*(3), 577–591. https://doi.org/10.1044/1092-4388(2005/040)

Suzukida, Y., & Saito, K. (2021). Which segmental features matter for successful L2 comprehensibility? Revisiting and generalizing the pedagogical value of the functional

load principle. *Language Teaching Research*, *25*(3), 431–450. https://doi.org/10.1177/1362168819858246

Thir, V. (2020). International intelligibility revisited: L2 realizations of NURSE and TRAP and functional load. *Journal of Second Language Pronunciation*, *6*(3), 458–482. https://doi.org/10.1075/jslp.20012.thi

Tribushinina, E., & Gillis, S. (2017). Advances and lacunas in usage-based studies of first language acquisition. In J. Evers-Vermeul & E. Tribushinina (Eds.), *Usage-based approaches to language acquisition and language teaching* (pp. 13–46). De Gruyter, Inc. http://ebookcentral.proquest.com/lib/aucegypt/detail.action?docID=4822114

Trofimovich, P., Isaacs, T., Kennedy, S., & Tsunemoto, A. (2022). Speech comprehensibility. In T. M. Derwing, M. J. Munro, & R. I. Thomson (Eds.), *The Routledge handbook of second language acquisition and speaking*. Routledge. https://doi.org/10.4324/9781003022497

Uchihara, T. (2022). Is it possible to measure word-level comprehensibility and accentedness as independent constructs of pronunciation knowledge? *Research Methods in Applied Linguistics*, *1*(2), Article 100011. https://doi.org/10.1016/j.rmal.2022.100011

Varonis, E. M., & Gass, S. M. (1982). The comprehensibility of non-native speech. *Studies in Second Language Acquisition*, *4*(2), 114–136. https://www.jstor.org/stable/44487230

Vilkaitė-Lozdienė, L., & Schmitt, N. (2019). Frequency as a guide for vocabulary usefulness: High-, mid-, and low-frequency words. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 81–96). Routledge. https://doi.org/10.4324/9780429291586

Wahba, K. M. (2021). Models of Arabic pronunciation. In K. C. Ryding & D. Wilmsen (Eds.), *The Cambridge handbook of Arabic linguistics* (pp. 127–152). Cambridge University Press. https://doi.org/10.1017/9781108277327.007

Wang, X. (2022). Segmental versus suprasegmental: Which one is more important to teach? *RELC Journal*, *53*(1), 194–202. https://doi.org/10.1177/0033688220925926

Weismer, G. (2008). Speech intelligibility. In M. J. Ball (Ed.), *The handbook of clinical linguistics* (pp. 568–582). Blackwell Pub. https://doi.org/10.1002/9781444301007.ch35

Yang, C., Chu, J., Chen, S., & Xu, Y. (2021). Effects of segments, intonation and rhythm on the perception of L2 accentedness and comprehensibility. In C. Yang (Ed.), *The acquisition of Chinese as a second language pronunciation: Segments and prosody* (pp. 233–255). Springer. https://doi.org/10.1007/978-981-15-3809-4_11

Zielinski, B. (2015). The segmental/suprasegmental debate. In M. Reed & J. M. Levis (Eds.), *The handbook of English pronunciation* (pp. 397–412). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118346952.ch22

Appendix A

IRB Approval

THE AMERICAN UNIVERSITY IN CAIRO
INSTITUTIONAL REVIEW BOARD

Case# 2022-2023-262

To: Mark Ferenc Papai
Raghda El Essawi
Sara Tarek
From: Heba Kotb
Chair of the IRB
Date 18/5/2023
Re: IRB approval

This is to inform you that I reviewed your revised research proposal entitled "**Segmental Correlates of L2 Arabic Speech Comprehensibility and Accentedness: L1 Consonant Age of Acquisition as a Potential Predictor**"

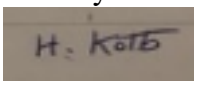
It required consultation with the IRB under the "expedited" category. As you are aware, there were minor revisions to the original proposal, but your new version addresses these concerns successfully. Your proposal used appropriate procedures to minimize risks to human subjects and that adequate provision was made for confidentiality and data anonymity of participants in any published record. I believe you will also make adequate provision for obtaining informed consent of the participants.

This approval letter was issued under the assumption that you have not started data collection for your research project. Any data collected before receiving this letter could not be used since this is a violation of the IRB policy.

Please note that IRB approval does not automatically ensure approval by CAPMAS, an Egyptian government agency responsible for approving some types of off-campus research. CAPMAS issues are handled at AUC by the office of the University Counsellor. The IRB is not in a position to offer any opinion on CAPMAS issues, and takes no responsibility for obtaining CAPMAS approval.

This approval is valid for only one year. In case you have not finished data collection within a year, you need to apply for an extension.

Thank you and good luck.



Heba Kotb
IRB chair, The American University in Cairo
2078 HUSS Building
T: 02-26151857
Institutional Review Board The American University in Cairo
AUC Avenue, P.O. Box 74
New Cairo 11835, Egypt.
tel 20.2.2615.1000
fax 20.2.27957565
Email: irb@aucegupt.edu
Email: hebakotb@aucegypt.edu

Appendix B

Raw Word List

Early acquisition	Late acquisition
حَميد	صَواب
حَلْبة	صارِم
حَفْنة	صَليب
حَسْرة	صَميم
حَراك	ظُلمة
حافِز	ضَجيج
حُزْمة	ضَباب
حُرْمة	خانة
عَتْمة	خُلود
عِنان	خَيْبة
عَنْوة	خَلَف
عِداد	خَراب
عِبْرة	غَمْرة
أُسْوة	غَفْلة
أَمَد	غَرام
هَيْبة	غارة
هاجِس	قُدوم
هُدْنة	قالِب
هَيْكَل	قامة
هِبة	قَبيل

Appendix C

Consent Form for Speakers

THE AMERICAN UNIVERSITY IN CAIRO
INSTITUTIONAL REVIEW BOARD

Documentation of Informed Consent for Participation in Research Study

You are being asked to participate in a research study. The purpose of the research is to investigate the perception of non-native Arabic speech, and the findings will be presented to a committee and may be published.

If you agree to participate in this study, you will be asked to do the following:

- Record yourself reading aloud a list of Arabic words
- Complete a questionnaire about your language experience

The expected duration of your participation is 15 minutes. There are no known risks or discomfort associated with this research.

Although you will receive no direct benefits, your participation in this research will help potentially improve the teaching of Arabic as a foreign language.

The information you provide for the purposes of this research is confidential.

Questions about the research, your rights as a research participant, or other research-related inquiries should be directed to Mark Papai at 01069842856, markpapai@aucegypt.edu.

Participation in this study is voluntary. Refusal to participate will involve no penalty or loss of benefits to which you are otherwise entitled. You may discontinue participation at any time without penalty or the loss of benefits to which you are otherwise entitled.

Signature _____

Printed Name _____

Date _____

Appendix D

Language Background Questionnaire for Speakers

1. Age
2. Gender
3. What is your highest level of education?
4. What is your first language (the language you spoke at home when growing up)?
5. What was your primary language of schooling? (leave blank if the same as first language)
6. At what age did you start learning Arabic?
7. How many years have you been learning Arabic?
8. How would you judge your current level of Arabic? (Novice/Intermediate/Advanced/Superior/Distinguished)
9. Have you lived in an Arabic-speaking country? List the country/countries, the time you spent in each one and your age/age range when living there.
10. What other foreign/second languages do you know? List your level for each. (Novice/Intermediate/Advanced/Superior/Distinguished)
11. Do you have experience in teaching a foreign language? If yes, list language and years of experience?
12. Have you received formal training in linguistics?
13. Have you received formal training in phonetics and/or phonology?
14. Have you received targeted instruction for Arabic pronunciation (e.g., tajwid, private classes, etc.)?

Appendix E

Consent Form for Listeners

استمارة موافقة مسبقة للمشاركة في دراسة بحثية

عنوان البحث : Segmental Correlates of L2 Arabic Speech Comprehensibility and Accentedness: L1 Consonant Age of Acquisition as a Potential Predictor

الباحث الرئيسي Mark Papai:طالب ماجستير
البريد الالكترونيmarkpapai@aucegypt.edu :
الهاتف01069842856 :

انت مدعو للمشاركة في دراسة بحثية عن الانطباعات السمعية لعناصر الكلام.

هدف الدراسة هو تحسين طرق تدريس اللغة العربية للناطقين بغيرها.

نتائج البحث ستنشر في موقع الجامعة الأمريكية بالقاهرة الإلكتروني .

المدة المتوقعة للمشاركة فى هذا البحث 15 دقيقة.

اجراءات الدراسة تشتمل على تقييم تسجيلات صوتية وإكمال استبيان قصير عن خلفيتك الديموغرافية واللغوية.

لا توجد لهذه الدراسة مخاطر متوقعة من المشاركة.

الاستفادة المتوقعة من المشاركة في البحث :ستفيد مشاركتك فهمنا للغة العربية وستساعد تحسين طرق تدريسها للناطقين بغيرها.

السرية واحترام الخصوصية :المعلومات التى ستدلى بها فى هذا البحث سوف تكون سرية.

أي أسئلة متعلقة بهذه الدراسة أو حقوق المشاركين فيها يجب أن توجه إلى الباحث الرئيسي Mark Papai ،01069842856، markpapai@aucegypt.edu.

ان المشاركة فى هذه الدراسة ماهى الا عمل تطوعى ,حيث أن الامتناع عن المشاركة لايتضمن أى عقوبات أو فقدان أى مزايا تحق لك. ويمكنك أيضا التوقف عن المشاركة فى أى وقت من دون عقوبة أو فقدان لهذه المزايا.

Appendix F

General Survey Instructions

من خلال هذا الاستبيان يُطلب منك الاستماع إلى **40** تسجيلا صوتيا ووضع تقييمات ردا عليها **.يرجى ارتداء سماعة كي تسمع التسجيلات جيدا** .يحتوي كل تسجيل على كلمة عربية واحدة فقط نطقها متحدثون أجانب ومصريون. تتضمن بعض التسجيلات أخطاء في نطق الحروف ف**عليك محاولة فهم الكلمة وتقييم صعوبة فهمها كما يُطلب منك تقييم مدى وجود لكنة أجنبية** لدى المتحدث حسب إحساسك. **لا توجد إجابات صحيحة أو خاطئة** إذ يستهدف هذا الاستبيان استخراج انطباعاتك الشخصية.

في الصفحتين التاليتين سيمكنك التمرين على التقييمات ثم ستقدم لك التسجيلات الحقيقية.

Appendix G

Stimuli Instructions and Rating Scales

استمع(ي) إلى الكلمة وقم/قومي بوضع التقييم

1. صعوبة الفهم ليست مفهومة على الإطلاق 1 _ _ _ _ _ _ _ _ _ 9 مفهومة تماما
2. مدى وجود لكنة أجنبية لا توجد لكنة أجنبية 1 _ _ _ _ _ _ _ _ _ 9 توجد لكنة أجنبية شديدة

Appendix H

Language Background Questionnaire for Listeners

1. السن
2. النوع
3. مكان النشأة
4. أعلى مستوى من التعليم
5. لغة التعليم
6. ما هي اللغات التي تعرف(ين)ها وما هو مستواك فيها؟
7. هل لديك خبرة في تدريس اللغات الأجنبية؟ اذكر(ي) اللغة وعدد سنوات خبرتك
8. هل لديك خبرة في البحث العلمي اللغوي؟
9. هل لديك خبرة في الصوتيات (مثلا التجويد)
10. هل لديك خبرة مع أجانب يتحدثون بالعربية؟