The Effects of English Connected Speech Processes on Adult Mandarin-Speaking Learners' L2 Speech Perception

Shuo-Fang Liang

St Edmund Hall

University of Oxford

Thesis submitted for the degree of

Doctor of Philosophy

Trinity Term 2025

Supervisors: Professor Elizabeth Wonnacott and Dr Robert Woore

Acknowledgements

This doctoral research has been subsidised by the 2024 National Science and Technology Council Taiwanese Overseas Pioneers Grants (TOP Grants) for PhD Candidates.

I am immensely grateful to my supervisors, Liz and Robert, without whom this PhD would not have been possible. Liz's incredibly sharp eye for methodological details saved me from making lots of technical mistake-*s* (emphasis on the plural). Robert is truly a caring mentor, who expanded my academic perspectives and nudged me towards the right mindset when I needed it most. I am extremely fortunate to have had Liz and Robert throughout this journey, with endless thanks and genuine appreciation.

I extend my sincere gratitude to Professor Sonali Nag, Professor Heath Rose, Dr Faidra Faitaki, and Professor Joan Mora for kindly reviewing my work at different stages. I also thank many course and lab mates who generously offered feedback, even when I shared ideas in their unpolished forms.

Lastly, to myself four years ago, thank you for plucking up the courage to embark on this adventure against all odds. Further ahead, there is still little known and much to strive for.

Abstract

Connected speech processes (CSPs) refer to alterations in pronunciation that occur when words are spoken in fluent speech, as opposed to when they are articulated individually and with care, known as their 'citation form'. For example, the two words *find* and *out* may sound different when each word is pronounced carefully in isolation, compared to when they are spoken spontaneously in a context like '*let's find out*' in which the phrase may sound remarkably similar to *fine doubt*. Research shows that these CSPs often impinge on second language (L2) learners' speech perception. However, few studies have systematically explored the relative impact of various CSP categories on learners with specific language backgrounds, and the extent to which learners may benefit from targeted pedagogical interventions in terms of intelligibility, comprehensibility, and general listening comprehension. Therefore, this research incorporates two cumulative phases, aiming to (1) identify the extent to which a range of English CSPs affect adult Mandarin-speaking learners' L2 speech perception; and (2) evaluate the effectiveness of distinct pedagogical interventions targeting the most challenging CSP categories for this learner population.

Phase 1: Diagnosis

In Phase 1, ten categories of English CSPs were selected for investigation. 50 adult native Mandarin-speaking participants with advanced proficiency in English were recruited as an experimental group, and 20 adult native English speakers served as a comparison group. Participants were assessed for their ability to transcribe words spoken in citation form versus the same words spoken in connected speech featuring the target CSPs. Results indicated that L2 participants experienced a significantly greater decline in transcription accuracy from citation form to connected speech, compared to L1

participants, suggesting that CSPs had a more detrimental impact on their performance. Critically, the level of difficulty varied across CSP categories, with disparities in accuracy between connected speech and citation form ranging from 13 to 49 percentage points. Six categories of CSPs were identified as particularly challenging for these L2 learners: *Deletion of Onset, Glide Insertion, Reduction of Stop, CC-V Linking, Palatalisation, Regressive Assimilation*.

Phase 2: Intervention

Based on the diagnostic results, Phase 2 implemented an intervention study to further examine the effects of teaching (the 'intervention') about the six CSP categories identified as the most challenging. 66 adult native Mandarin-speaking university students from Taiwan studying English-related disciplines and possessing upper-intermediate to advanced proficiency were recruited. Participants were randomly assigned to one of three groups, each receiving distinct pedagogical interventions. The first group received explicit rule explanation of CSPs, the second group was guided to notice pronunciation alterations without explicit rule explanation, and the third, a control group, engaged in listening comprehension practice without a direct focus on pronunciation. All participants completed a pre-test, an intervention programme, and a post-test. Language input throughout the interventions was highly controlled across all three groups, despite the different pedagogical approaches: sample words and sentences illustrating the target CSPs were identical in all three conditions.

Bayesian mixed effects regression modelling and Bayes factor analysis were employed to evaluate the effectiveness of the interventions. Results indicated an overall improvement in intelligibility, as measured by transcription accuracy of CSP-affected

words. However, gains in general listening comprehension, though numerically present, were not statistically robust. Unexpectedly, participants exhibited a decline in comprehensibility ratings, based on their perceived ease of understanding the connected speech items. Despite the main effects observed when data across all groups were aggregated, there was no evidence of interactions between group (three conditions) and assessment (pre-test and post-test) for these measures, suggesting that the relative effectiveness of the three intervention conditions remained inconclusive: neither the alternative hypothesis (H_1: differential effects across conditions) nor the null hypothesis (H_0: no differential effects across conditions) was supported by the data. The findings are discussed in relation to their theoretical and pedagogical implications, along with directions for future research on CSPs and L2 speech perception.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

BF	Bayes Factor
CSPs	Connected Speech Processes
EFL	English as a Foreign Language
ESL	English as a Second Language
GLC	General Listening Comprehension (L2)
L1	First Language
L2	Second Language
LexTALE	Lexical Test for Advanced Learners of English
LSK	Lexical-Semantic Knowledge Test (L1)
Ofsted	Office for Standards in Education, Children's Services and Skills (government department which inspects educational services in England)

Chapter 1: Introduction

1.1 Research Background

Comprehending natural spoken language in real time is a complex and cognitively demanding task. Unlike written texts, speech is inherently continuous, transient, and devoid of clear markers of word boundaries. Listeners must rapidly decode incoming acoustic signals and transform them into meaningful units, a process that engages a series of intricate cognitive mechanisms informed by multiple domains of linguistic knowledge (Lynch, 1998; Martínez-Flor & Usó-Juan, 2006; Vandergrift & Goh, 2012). While this process typically becomes automatic and intuitive for first language (L1) listeners through prolonged immersive language exposure and verbal interaction, it is, however, often a challenging task for second language (L2) listeners, particularly when confronted with streams of continuous speech (Field, 2003).

One pervasive barrier for L2 listeners is their difficulty in processing phonological alterations that characterise spontaneous speech, which deviate from clearly enunciated forms of pronunciation (Shockey, 2003). This linguistic phenomenon, known as connected speech processes (CSPs) (Alameen & Levis, 2015), encompasses a variety of categories such as linking, insertion, assimilation, deletion, and reduction of sounds. From the speakers' perspective, these alterations serve to facilitate fluidity and economy of articulation in naturally produced speech (Ladefoged & Johnson, 2015). For example, consider how the two words *find* and *out* may sound different when each word is pronounced individually and with care, compared to when it is spoken spontaneously in a context like '*let's find out*'. In a natural, fluent utterance – or technically, connected speech – this phrase may sound similar to *fine doubt*. Such differences can substantially hinder communication, particularly between L1 and L2 interlocutors, when the latter

struggle to recognise words affected by discourse features which tend to occur frequently and systematically in the speech of the former (Bowen, 1975; Gilbert, 1995).

For L2 listeners, CSPs often obscure lexical forms, complicating the task of mapping incoming acoustic signals to known words. These altered forms (induced by CSPs) contrast in various ways (see Section 2.1 for a review of CSP categories and Section 4.1.2 for the categories investigated) with the 'canonical' pronunciation of words, also known as their citation forms (i.e. when individual words are spoken carefully in isolation, as commonly found in dictionaries). However, it is these citation forms that are often taught in L2 learning contexts (Brown, 2013). As a result, many L2 learners may lack the (implicit or explicit) knowledge of CSPs or the awareness of these CSPs in connected speech. This not only slows down L2 listeners' speech processing, but often leads to frequent failure in recognising words which they otherwise might be able to recognise (Vidal, 2019).

While much of the variability in CSP patterns tends to follow predictable phonological rules (Hieke, 1987a), their perception by L2 learners' often remains elusive without pedagogical support. In the absence of targeted instruction, learners are left to develop coping strategies on their own to process speech abounding with CSPs, often with limited success (e.g. Ahmadian & Matour, 2014). Thus, it is not uncommon for L2 learners to master advanced grammar and vocabulary but still experience substantial difficulties in understanding connected speech, particularly when produced by native speakers in authentic communicative contexts (e.g. Brown, 2013).

In recent decades, these perceptual challenges inherent in L2 speech processing have

gained growing attention in the field of second language acquisition. One critical issue identified is that the insufficient development of L2 listening ability is often associated with exposure to predominantly non-authentic, over-articulated, or foreigner-oriented input (e.g. Cruttenden, 2014; Jones & Ono, 2001). Although there has been increased recognition of the importance of understanding connected speech, teaching practitioners continue to grapple with several practical concerns. These range from the lack of systematic instruction in CSPs within available mainstream materials, through limited empirical evidence to guide pedagogical decisions, to inadequate teacher training in the instruction of listening and CSPs (e.g. Ito, 2006; Crawford & Ueyama, 2011). Consequently, many teachers may lack sufficient awareness, theoretical knowledge, pedagogical resources, and suitable instructional strategies to address CSPs effectively in the classroom.

Perhaps most pressing is a paucity of robust empirical evidence pinpointing which categories of CSPs impose greater challenges for specific L2 learner populations and should therefore be given priority for pedagogical treatment. In other words, there is yet to be a consensus on which CSP categories are most challenging for particular groups of L2 listeners. Many existing teaching materials and practices thus often rely on intuition or broad generalisations of CSPs rather than research-informed decisions. Furthermore, existing evidence on the effectiveness of different pedagogical approaches for teaching CSPs remains inconclusive.

Given these challenges, there is a need for more comprehensive empirical studies on CSPs, as well as on instructional practices that can effectively support L2 learners. The issues outlined above also underscore the importance of situating CSPs within the broader

context of L2 listening development. This research seeks to address these gaps by investigating the perceptual impact posed by CSPs and evaluating the effectiveness of distinct pedagogical interventions targeting particularly challenging CSP categories. The following section outlines the specific research aims of this inquiry.

1.2 Research Aims

This research comprises two distinct yet cumulative phases, each designed to address a specific research aim concerning the effects of CSPs on L2 speech perception. The first phase investigates the extent to which a range of English CSPs impact adult Mandarin-speaking learners' L2 speech perception. The second phase evaluates the effectiveness of different pedagogical approaches for teaching CSPs to this learner population. The research questions guiding this inquiry are presented in Section 2.5.

Phase 1: Diagnosis

This diagnostic phase aims to identify which categories of English CSPs impose greater or lesser degrees of difficulty on L2 learners' ability to decode connected speech. By examining patterns of perceptual difficulty, the study seeks to determine which CSP categories are particularly problematic for the target learner population. Crucially, findings from this phase serve as an empirical basis to inform the contents of the programmes of instruction which will be evaluated in the subsequent phase of the study.

Phase 2: Intervention

Building on the diagnostic findings from Phase 1, the second phase aims to evaluate the effectiveness of teaching English CSPs (the 'intervention') to the target learner population. This phase implements pedagogical interventions addressing the CSP categories

identified as most challenging in Phase 1, and specifically compares three distinct pedagogical approaches in terms of their effectiveness in enhancing learners' L2 speech perception. The target CSP categories and intervention conditions are detailed in Section 5.1. Findings from this phase are intended to provide empirically grounded insights into effective strategies for incorporating CSPs into L2 teaching practice.

1.3 Thesis Overview

This thesis is structured in seven chapters. Following this introduction:

Chapter 2: Literature Review

This chapter reviews literature across several areas pertinent to this research: the nature of CSPs, theories of speech perception, and the teaching and learning of listening and CSPs in L2 contexts. It begins by defining CSPs and reviewing terminology, categorisation, and phonological mechanisms, as well as their manifestation in spoken English. Subsequently, the chapter delineates the scope of speech perception in this research, explores relevant theoretical models and hypotheses, and attends to the interplay of top-down and bottom-up processing. The review then turns to the pedagogical treatment of listening and CSPs in language education, underscoring their importance in L2 learning and their relative neglect in teaching practice. It further provides a critical synthesis of empirical studies on CSPs, along with an account of methodological challenges and research gaps. The chapter concludes with a summary of the literature review that establishes a rationale for the current research, culminating in the formulation of research questions guiding this inquiry.

Chapter 3: General Methodology

This chapter outlines the overarching methodological framework of the research,

describing the quantitative and experimental design adopted across the two cumulative phases. It introduces the core constructs of intelligibility and comprehensibility, which underpin the outcome measures in both diagnostic and intervention studies. It also elaborates on the operationalisation of individual linguistic measures, along with their implementation across both phases. This chapter establishes the conceptual and procedural foundation for the empirical investigations that follow.

Chapter 4: Phase 1 Study – Diagnosis

This chapter presents the methods and findings of the diagnostic study. The methods section details participant profiles, the selection of CSP categories, the development of the stimulus inventory, and implementation procedures. The results section reports statistical analyses that reveal L2 participants' substantial difficulty in recognising CSP-affected words, as well as varying degrees of impact across different CSP categories. Based on the empirical data, quantitative thresholds are established to determine which categories are particularly challenging and thus warrant prioritisation for pedagogical interventions. Supplementary measures (L1 lexical-semantic knowledge, L2 general listening comprehension, and LexTALE) are also examined in terms of their distributions and correlations with principal measures. The chapter concludes with a summary of the study and reiterates the target CSP categories subject to the subsequent intervention study.

Chapter 5: Phase 2 Study – Intervention

This chapter reports on the intervention study, which evaluates the effectiveness of pedagogical interventions targeting the CSPs identified as particularly challenging in Phase 1. The methods section outlines participant profiles, the selected CSP categories and stimuli, implementation procedures, and the pedagogical approaches employed. The

results section presents findings for both intelligibility and comprehensibility measures, analysed using Bayesian mixed effects regression models and Bayes factors. Additionally, general listening compression is examined to explore broader perceptual development beyond processing CSP-affected items. Correlation analyses between principal and supplementary measures are also conducted to discern their interrelationships before and after the interventions. The chapter concludes with a summary of major findings, leading into interpretations in the subsequent Discussion chapter.

Chapter 6: Discussion

This chapter synthesises and interprets the findings from both empirical phases, structured around the research questions. Each section addresses a specific research question and its associated sub-questions, situating the findings withing the broader context of existing literature. The discussion unfolds sequentially: it begins with the impact of different CSP categories, followed by the effectiveness of the interventions on distinct linguistic measures, and finally addresses the observed correlations among those measures. The latter part of the chapter considers practical implications for teaching CSPs, including practitioners' metalinguistic awareness, the role of input variety and authenticity, and the applicability of CSP instruction across diverse pedagogical contexts. Overall, this chapter intends to translate empirical findings into pedagogically meaningful insights for L2 teaching practice.

Chapter 7: Conclusions

This chapter summarises the key findings of the research and articulates its theoretical and practical contributions. Acknowledging this work's limitations, it also proposes directions for future inquiry. Finally, the thesis concludes with reflective closing remarks.

Chapter 2: Literature Review

2.1 Connected Speech Processes (CSPs)

To provide a clear context for this research, this section presents an overview of connected speech processes (CSPs). It first reviews the terminology and definition of CSPs, and then introduces the phonological mechanisms along with surface realisations characteristic of various CSP categories in spoken English.

2.1.1 Terminology and Definition

Connected speech processes (CSPs) are defined as 'the differences from citation pronunciations that occur when words occur in normal spoken discourse' (Alameen & Levis, 2015, p. 1). Various terms have been used interchangeably to denote these phonological alterations, including Sandhi-variation (e.g. Henrichsen, 1984), reduced forms (e.g. Brown & Hilferty, 1986), dynamic speech (e.g. Hieke, 1987b), absorption (e.g. Hieke, 1987a), relaxed speech (e.g. Weinstein, 2001), and casual speech (e.g. Shockey, 2003). While the current research adopts the term 'CSPs', some alternatives such as 'Sandhi-variation' or 'reduced forms' may be referred to when reviewing studies in which those terms are used.

CSPs can be understood by contrasting two linguistic concepts: citation form and connected speech. The citation form represents the pronunciation of individual words when they are articulated carefully in isolation, often conforming to conventional dictionary pronunciations. Conversely, connected speech refers to spontaneous natural spoken discourse in which individual words are pronounced with less precision in combination with other words, resulting in deviations from their citation form. The phonological processes that transform the citation form into connected speech are termed

CSPs.

This linguistic phenomenon is rooted in the principle of articulatory economy, whereby speakers tend to convey meaning with minimal effort (Ladefoged & Johnson, 2015). In other words, the guiding principle in connected speech is the 'ease of articulation' (ibid., p. 251), prompting speakers to link, insert, modify, delete, or reduce certain sounds during speech. Moreover, when multiple CSPs co-occur within a phrase, the chaining of multiple rules can lead to substantial changes in surface word forms (Darcy et al., 2009). The following section reviews the phonological mechanisms of common English CSPs, supplemented with phonetic details necessary for clarity.

2.1.2 Phonological Mechanisms of Common CSPs in English

As mentioned, the terminology of CSPs varies widely, and so does the taxonomy. Therefore, in addressing various categories of CSPs, it is considered more effective to base the discussion on one classificatory system. The review below follows primarily the classification proposed by Alameen and Levis (2015). Nevertheless, insights from alternative schemes provided by Celce-Murcia et al. (2010) and Hieke (1987a) are referenced, along with relevant literature where appropriate.

Linking

Linking emerges as the simplest (Alameen & Levis, 2015), mildest (Hieke, 1987a), and most frequently used (Celce-Murcia et al. 2010) CSP in English. This CSP category, which encompasses various subcategories depending on the segmental features concerned, fundamentally aims to 'avoid hiatus during phonation' (Hieke, 1987a, p. 43). One type of linking occurs when a word-final vowel-consonant (VC) sequence is

followed by a word-initial vowel (V), indicated as VC-V linking (Celce-Murcia et al., 2010). For example, the consonant /s/ in *face it* tends to become ambisyllabic, serving as the coda of *face* and the onset of *it* simultaneously, and is realised as /feɪs͡ɪt/[1]. This linking of the /s/ sound shortens the temporal gap between *face* and *it*, eliminating an audible pause that would otherwise mark the word boundary, as a physical space does in the written form.

A similar yet slightly different type of linking involves a CC-V sequence, in which the last segment of a word-final two-consonant cluster becomes the onset of the following vowel-initial word, known as resyllabification (Celce-Murcia et al., 2010). For example, in the phrase *find out*, the consonant /d/ becomes the onset of *out*, resulting in the pronunciation of /faɪn.daʊt/, which would sound similar to *fine doubt*.

Additionally, linking can occur with a C-C sequence when a word-final consonant is identical to the initial consonant of the following word, in which case the sound is elongated instead of being pronounced twice separately, a phenomenon referred to as gemination[2]. For example, *save videos* is usually realised as /seɪvːɪdiəʊz/ in connected speech. While geminate consonants are sometimes classified as a form of elision (e.g. Liang, 2015), their lengthening property distinguishes them as a case of C-C linking in the current research.

[1]A ligature tie and an under tie on the two sides of the segment /s/ indicate its ambisyllabic characteristic.

[2]Note that Hieke's (1987a) refers to this process as 'degemination', describing a phenomenon in which a consonant is not articulated twice but instead realised as a single segment. The prefix 'de-' denotes the avoidance of repetition (i.e. 'gemination' in its sense of doubling). However, as the emphasis here is on the lengthening resulting from single articulation, the term 'gemination' is adopted instead, as indicated by the lengthening diacritic (ː) marking the geminate consonants. For a discussion of different types of degemination, including cases without lengthening, see Hieke (1987a).

Notably, linking is so prevalent that in an early study by Hieke (1984), it is considered 'a regular rule of English', which accentuates its indispensability to fluency in the language (p. 351). More broadly, because linking serves a crucial function in connecting sounds to create a smooth flow of speech, it is sometimes viewed as the fundamental purpose of CSPs, implying that other categories of CSPs facilitate the linking process through various alternative mechanisms, which will be explored below.

Insertion

In connected speech, additional segments may be inserted that are not present in the citation form. One common type of insertion involves the epenthesis of glides between two vowels. Specifically, the glide /j/ is inserted when the preceding vowel is a close front /i/ or /ɪ/, while /w/ is inserted when the first vowel is a close back /u/ or /ʊ/. For example, *see it* can be realised as /siːjɪt/, and *low income* can be realised as /ləʊwɪŋkʌm/. Interestingly, some studies (e.g. Anderson-Hsieh et al., 1994) classify this process as a type of linking for V-V sequences, an approach also reflected in Hieke's (1987a) classification, which labels it 'glide attraction'. The current research, however, in line with Alameen and Levis's (2015) classification, treats these glides as a process of insertion rather than linking, because of their epenthetic nature as opposed to the existing segments in the citation form. Additional types of insertion include the so-called 'intrusive *r*', which is more prevalent in British English varieties in cases like *idea-*/ɹ/*-of* (Alameen & Levis, 2015), and word-internal insertion in cases like /t/ in *prin(t)ce* and /p/ in *com(p)fort*. The latter may not necessarily be considered CSPs, as they can affect words even in isolation (see Section 4.1.2: *Criterion 1: Inter-Word Phonological Environment*).

Modification

Modification involves changes in specific phonological features of segments. This category includes processes such as palatalisation and assimilation. Palatalisation occurs when alveolar fricatives /s/, /z/, stops /t/, /d/, or their combinations /ts/, /dz/ are followed by the glide /j/. In these sequences, they can be realised together as post-alveolar fricatives /ʃ/, /ʒ/ or affricates /tʃ/, /dʒ/. Palatalisation is particularly common when the second lexical item is a second person pronominal form such as *you* or *your*. For example, *makes you* can be realised as /meɪkʃu/ and *got your* as /gɒtʃʊr/.

Regressive assimilation occurs when a sound changes to align more closely with a following segment. An example is the alveolar nasal /n/ being assimilated to the following labial or velar consonant in place of articulation while preserving the nasal property, hence realised as /m/ or /ŋ/, respectively. For example, *green box* can be realised as /griːmbɒks/ and *thin glass* as /θɪŋglɑs/. In contrast, when assimilation occurs in the opposite direction – i.e. where a preceding sound causes a following sound to change – it is known as progressive assimilation. As noted by Celce-Murcia et al. (2010), 'regressive assimilation is more pervasive as a purely phonological process' (p. 168), whereas progressive assimilation is often related to morphophonological contexts. A typical case of the latter is the plural morpheme *-s* being realised as /s/ in *cats* and /z/ in *dogs*, depending on the voicing status of the preceding segment, which again is not necessarily a CSP as it affects words even in citation form.

Alameen and Levis (2015) also include two additional types of modification in their classification: flapping, which is more characteristic of American English (e.g. an intervocalic /t/ realised as a flap /ɾ/ in *get around*); and glottalisation, which is more

characteristic of British English (e.g. a word-final /t/ realised as a glottal stop /ʔ/ *in fat man*).

Deletion

Deletion refers to complete omission of segments in connected speech. This process can occur, for example, at the onset of pronouns and auxiliaries, or within consonant clusters. The former involves what is referred to as 'breathed onset deletion' (Hieke, 1987a), where the breathed glottal fricative /h/ at the beginning of unstressed pronominal forms (e.g. *he, his, him, her, hers*) or auxiliaries (e.g. *have, has, had*) can be deleted, especially when these items do not appear in an utterance-initial position. For example, *watch him* can be realised as /wɒt͡ʃɪm/. A similar deletion also occurs with inter-dental fricative /ð/, as in *tell them*, which can be realised as /tɛləm/.

The other type involves the omission of a medial consonant in a cross-word three-consonant cluster (i.e. CC-C sequences). In such contexts, the alveolar /t/ and /d/ are particularly prone to deletion. For example, *last present* can be realised as /lɑspɹɛzənt/. The segments (/t/ and /d/) are typical targets for deletion, not only because they frequently appear at the ending position within coda clusters due to phonotactics, but also because English coronals can be considered underspecified[3] (Avery & Rice, 1989) and thus may be more susceptible to phonological variation.

[3]Based on the phonetic inventory of English, including examples such as allophonic alternation between /t/ and /ʔ/ and the nasal assimilation of /n/ to adjacent consonants, Avery and Rice (1989) argue that coronals are phonologically underspecified in English. This means coronals are relatively unmarked compared to labial and dorsal consonants, and thus their place specification (Coronal node) is absent in the underlying phonological representation. Instead, corornals can be represented more abstractly by its hypernymic feature, a Place node. For a detailed theoretical grounding of these assumptions, including a discussion of Node Activation Condition (NAC), see the original paper (Avery & Rice, 1989).

Additionally, Alameen and Levis (2015) include certain contractions (e.g. *can't* for *can not*) under the category of deletion, while Celce-Murcia et al. (2010) note a few word-internal types of deletion such as syncope (e.g. *camera* realised as /ˈkæm.rə/), aphesis (e.g. *about* realised as *'bout*), and loss of /ɹ/ (e.g. *Febɹuary* realised as /ˈfɛbjʊri/). These types of deletion can be regarded as lexicalised to some extent, in that they are usually associated with specific lexical items rather than purely phonological contexts.

Reduction

Reduction can apply to both vowels and consonants in connected speech. In English, vowel reduction is often associated with unstressed syllables, particularly in function words. For instance, *to* is frequently realised with schwa as /tə/ and *of* as /əv/, when unstressed. These realisations are sometimes referred to as weak forms (e.g. Matsuzawa, 2006; Kul, 2016), reflecting the 'weakened' (or 'neutralised') properties relative to their stressed counterparts.

Consonant reduction, as outlined in Alameen and Levis (2015), is observed in the unreleased stops at the word-final position when a following word begins with either a stop or an affricate. Noteworthily, in these contexts, while the air release of the stop is reduced, the articulatory closure of the stop (during which airflow is momentarily halted) is maintained. For example, *bad guy* can be realised as /badˀgaɪ/[4] and *blog post* as /blɒgˀpəʊst/, with the final /d/ and /g/ respectively produced without audible release.

Celce-Murcia et al. (2010), however, classify such consonant reduction either as a type

[4]The diacritic ˀ indicates 'no audible release' (International Phonetic Association, 2020).

of linking for [stop + stop] or [stop + affricate] sequences, or as a type of assimilation. While the current research agrees with the characterisation of these phonological conditioning environments – namely, [stop + stop] and [stop + affricate] – it contends that unreleased stops fall within the category of reduction, aligning with Alameen and Levis (2015) and reserving the category of linking for processes where the phonetic properties of segments involved remain intact.

Multiple

This category encompasses several high-frequency and, to some degree, lexicalised phrasal expressions that are affected by multiple CSPs simultaneously. For example, *gonna* is an informal yet conventional orthographic form of *going to*, which involves at least nasal assimilation, deletion of /t/, and vowel reduction. In contrast, **useta*[5] is an informal and non-standard orthographic form of *used to*, which involves medial consonant deletion, assimilation of non-voicing, and vowel reduction. Other instances including **whaddaya* (from *what do you* or *what are you*), **hafta*/*hasta* (from *have*/*has to*), and **supposta* (from *supposed to*) are documented in Weinstein (2001).

Additionally, certain contractions that involve more than the deletion of individual segments, such as *won't* (from *will not*) and *it's* (from *it is* or *it has*), may also be considered within this category. As noted by Alameen and Levis (2015), these forms undergo multiple phonological changes, representing outcomes of different CSPs rather than instances of single-process categories.

2.1.3 Summary

[5]An asterisk mark indicates non-conventional orthography.

This section has reviewed the definition of, and outlined various categories of, CSPs based primarily on the classification proposed by Alameen and Levis (2015). The review has illustrated that different CSPs vary considerably in how they alter the canonical pronunciation (i.e. citation form) of words, with each category inducing unique phonological mechanisms. The rationale and criteria for selecting CSP categories for empirical investigation – aligned with the classification adopted herein – are detailed in Section 4.1.2.

2.2 Speech Perception

This section explores the conceptual framework of speech perception in relation to the current research context, and reviews key models and hypotheses, including those that explicitly address the perception of connected speech in L2 contexts.

2.2.1 Definition and Scope

According to Strange and Shafer (2008), 'perception is, by definition, an internal mental (and psychological) process by which the perceiver recognizes incoming stimulus events as instances of mental categories' (p. 159). This suggests that during a speech event, the perceiver needs to first differentiate language sounds from other acoustic-phonetic signals, and subsequently map these language sounds onto established phonetic categories.

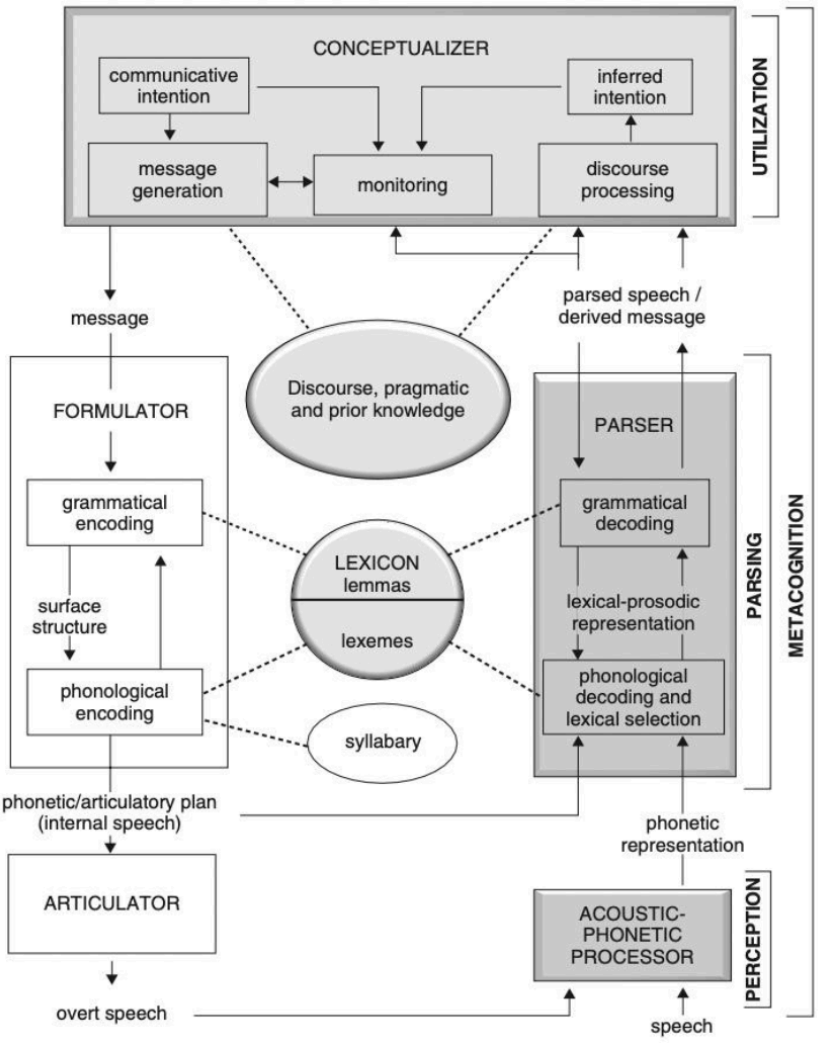
The current research adopts an inclusive definition of *speech perception* which overlaps with a related construct, *spoken word recognition*. Traditionally, research on spoken word recognition focuses on how phonetic and phonological cues (or other linguistic information) activate the listener's mental lexicon for identifying specific words (e.g. Marslen-Wilson & Welsh, 1978; Morton, 1969). However, if viewed from a broader

perspective, both speech perception and word recognition can be considered in service of comprehension. The general distinction between the two, though contentious and barely clear-cut, can be made by conceptualising different stages of speech processing.

In the Speech Production and Comprehension Model proposed by Vandergrift & Goh (2012, based on Levelt, 1993), three stages of processing – perception, parsing, utilisation (Anderson, 1995) – constitute the pathway to comprehension (Figure 2.1). Within this framework, *perception* corresponds to a fundamental stage that involves the reception of auditory signals and formation of initial phonetic representations, while word recognition emerges in the *parsing* stage, through processes such as phonological decoding and lexical selection.

Figure 2.1

Speech Production and Comprehension Model (Vandergrift & Goh, 2012)



Note. This figure, retrieved from Vandergrift and Goh (2012, p. 39), was developed based on the schematic representation by Levelt (1993), as reported by the authors.

Despite the conceptual overlap between these processes, the current research explores the nexus of L2 learners' perception of connected speech and their ability to recognise lexical items affected by CSPs. The term *perception* is therefore adopted in a broader sense, encompassing the transformation of acoustic signals into phonetic representations, followed by phonological decoding and lexical selection. The rationales for this broader interpretation are threefold:

1. **Theoretical Rationale:** According to Vandergrift and Goh's framework (ibid.), phonetic representations formed during the perception stage are immediately submitted to downstream processes of phonological decoding and lexical selection. Thus, it is conceptually more coherent to frame perception as encompassing these closely interlinked processes in parsing, rather than artificially isolating them or reversing the direction of their relationship.

2. **Methodological Rationale**: Traditional studies on spoken word recognition typically utilise tasks such as gating, shadowing, or error detection. Adopting a broader interpretation of *perception* aligns the scope of this study with its specific focus on how CSPs impact the intelligibility and comprehensibility (see Section 3.2 for the distinction between these two constructs) of connected speech for L2 learners.

3. **Task-based Rationale:** The principal measures in this study require participants to perceive speech sounds to complete dictation and comprehensibility rating tasks (see Section 3.3.1 for task descriptions). While dictation necessarily entails perception, it primarily taps into word recognition as the final outcome to be assessed, whereas the comprehensibility rating task specifically captures L2 learners' perceptual effort in understanding connected speech (independent of their success in word recognition). Together, these tasks inherently involve both perception (as in Figure 2.1) and word recognition, with the latter conceptualised as part of parsing in Vandergrift and Goh's model, particularly in relation to the dictation tasks.

For these reasons, *speech perception* in this research is operationalised through both connected speech dictation and comprehensibility rating tasks. These assessment

measures are elaborated in detail in Section 3.3. The following section reviews key models and hypotheses of speech perception in relation to the scope of this research.

2.2.2 Models and Hypotheses

In recent decades, L2 speech perception has appeared to attract burgeoning research interest. A number of models and hypotheses have been proposed to theorise how L2 listeners process speech differently than L1 listeners do. This section outlines four of those that are considered representative and relevant to the context of the current study: the Speech Learning Model, the Linguistic Perception Model, a phonological compensation hypothesis, and a phonological correlate model of reduced forms perception.

It is acknowledged that many other models have also contributed substantially to the field, including the Perceptual Assimilation Model (Best, 1995), the Phonological Interference Model (Brown, 1998), the Native Language Magnet Model (Kuhl, 1991; 2000), the Ontogeny Phylogeny Model (Major, 2001; 2002), the Polysystemic Speech Understanding Model (Hawkins & Smith, 2001). However, these models may be less relevant to CSPs, may not explicitly address mechanisms of L2 phonology development, or may bear similar assumptions. Therefore, they will not be reviewed individually within the scope of this study.

Speech Learning Model

The Speech Learning Model (SLM), originally proposed by Flege (1995, 2003) and revised as SLM-r (Flege & Bohn, 2021), posits that L1 and L2 phonetic categories coexist within a common phonological space. According to the SLM-(r), the formation of a new

L2 sound category depends on several phonetic factors, including the precision of the learner's L1 categories at the onset of L2 exposure, the quantity and quality of L2 input received, and critically, the phonetic dissimilarity between the L2 sound and the closest L1 category. Of these, the sound dissimilarity tenet is particularly influential and has been extensively examined experimentally. Specifically, the more an L2 sound is 'different' (i.e. perceptually distant) from the nearest L1 sound, the more likely learners will form a distinct new phonetic category for it. Conversely, if an L2 sound is perceptually too similar to an existing L1 category, it will likely be assimilated into that category (as opposed to being established as a separate category), potentially impeding L2 perception and production.

Early empirical support informing this SLM tenet was provided by Flege (1987), who demonstrated that the production of the French vowel /y/ by English learners of French was closer to the pronunciation of this vowel by L1 French speakers (in terms of F2 frequencies) than their production of French vowel /u/. This was likely attributable to the greater phonetic distance between the French /y/ (a 'new' phone for English speakers) from the English vowel /u/, compared to the French /u/, which closely resembled the English /u/. A related line of evidence was supplied by Flege et al. (1997), who found that speakers of German, Spanish, Mandarin, and Korean with greater English-language experience produced and perceived English vowels more accurately than their less experienced counterparts. Crucially, their results emphasised that the extent of this improvement depended on the structural (dis)similarity between the L1 and L2 vowel systems. For instance, German participants, whose L1 phonetic system includes a similar /i/-/ɪ/ phonemic contrast as in English, outperformed Spanish participants, whose L1 does not share this distinction. This advantage arose likely because German learners may have

readily perceived these English vowels as phonetically dissimilar, whereas Spanish speakers appeared to assimilate the two vowels into a single category, perceiving them as too similar to establish a robust contrast. Notably, no such advantage was observed for the English /ɛ/-/æ/ contrast, which is absent from both German and Spanish vowel inventories. These findings offer empirical support for the hypothesis that perceived dissimilarity of L2 realisations serves as an important determinant in the formation of new L2 categories.

Cross-language mapping studies (e.g. Iverson & Evans, 2007; 2009) provide further insights into the role of L1 phonetic systems in L2 vowel perception. In their 2007 study, Iverson and Evans examined English vowel perception among speakers of Spanish, French, German, and Norwegian, whose L1 vowel systems vary in size and complexity. Although learners with larger vowel inventories (German and Norwegian) outperformed those with smaller inventories (Spanish and French) in overall identification accuracy, all groups relied on similar acoustic cues such as formant movement and duration. Furthermore, perceptual vowel space mapping suggested that these learners appeared to form distinct L2-specific categories, rather than merely assimilating English vowels into pre-existing L1 categories. This aligns with the SLM's proposition that perceptual differentiation between L1 and L2 categories is critical for new category formation, although the process may initially be influenced by the structure of the learner's L1 phonetic system.

In a follow-up auditory training study, Iverson and Evans (2009) explored the impact of L1 category structure on L2 vowel learning by comparing Spanish and German learners of English, who underwent high-variability phonetic training on English vowels.

Interestingly, although German learners demonstrated greater immediate gains after five training sessions – despite their more densely packed L1 vowel space – Spanish learners eventually achieved comparable improvement following ten additional sessions. These findings lend partial support to the SLM by indicating that while initial stages of L2 perception are influenced by L1 categories, the development of new sound categories is possible across diverse L1 systems (e.g. with small or big vowel inventories). However, the results also appeared to complicate the SLM's expectation that learners with more densely populated vowel spaces (e.g. German speakers) would face greater difficulty forming new L2 categories due to *less* distance between vowels – though, as the authors pointed out, it remained questionable whether their more rapid progress involved forming genuinely new categories.

Levy's (2009a, 2009b) studies also contribute insights relevant to the assumptions regarding the influence of L1 phonetic categories on L2 perception. In a perceptual assimilation task, Levy (2009a) observed that American English learners of French, regardless of their level of language experience, frequently assimilated the French front rounded vowels /y/ and /œ/ onto English back vowels such as /u/ and /ʊ/, particularly in alveolar consonantal contexts (e.g. /radVt/[6]). However, assimilation patterns varied with language experience and phonological contexts, with learners who had extensive immersion experiences *less* frequently assimilating /y/ to /ju/ in bilabial contexts (e.g. /rabVp/). In a subsequent vowel discrimination experiment, Levy (2009b) demonstrated that increased language experience was associated with improved accuracy in

[6]These syllables (e.g. /radVt/ and /rabVp/) are pseudowords in which 'V' denotes the target vowel; they were used in controlled carrier phrases to systematically test vowel perception in specific phonological contexts (see Levy, 2009a, 2009b).

distinguishing French vowel contrasts, though certain contrasts, notably /y-u/, remained persistently difficult even for advanced learners. Importantly, participants' ability to discriminate vowel contrasts was related to their patterns of perceptual assimilation, highlighting the influence of perceptual differentiation on the development of new phonetic categories. These findings are broadly consistent with the SLM's proposition that the formation of new L2 categories depends on learners' sensitivity to cross-language phonetic differences and can be facilitated by extensive input and experience.

In sum, the SLM provides foundational predictions for language development by emphasising how L1-shaped segmental categories may constrain L2 speech perception and production. Nevertheless, it does not seem to fully explicate the cognitive mechanisms through which phonetic categories can be internalised (although it does acknowledge continuous interaction between L1 and L2 systems, cf. Flege & Bohn, 2021). Critically, the model is primarily concerned with segmental distinctions and does not address higher-order linguistic units beyond the phoneme level (Flege, 1995), such as those encountered in spontaneous connected speech. Therefore, to adequately account for the dynamic and variable nature of CSPs, as well as the related challenges faced by L2 learners, complementary theoretical frameworks capable of dealing with nuanced speech patterns beyond the segmental level may be required.

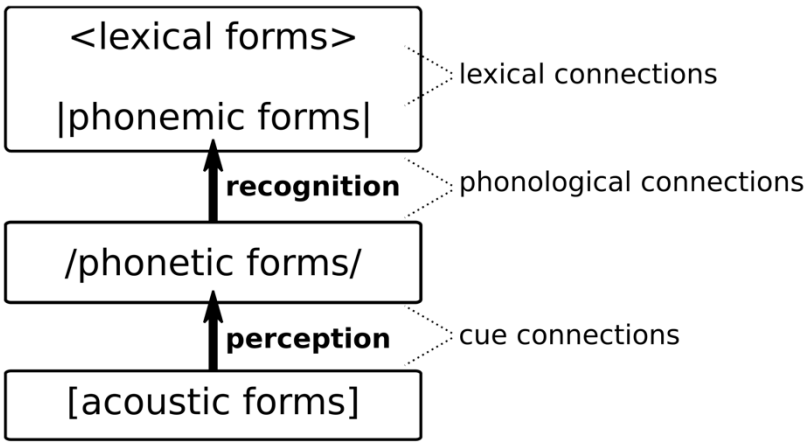
Linguistic Perception Model

The Linguistic Perception (LP) Model (Escudero, 2005) conceptualises *perception* as a process that transforms auditory input into perceptual input, which is then processed by a *recognition* system. A crucial component during perception is a linguistic processor termed *perception grammar*, which uses both phonetic constraints and language-specific

phonology to map the acoustic signal onto perceptual categories of speech sounds. These perceptual forms subsequently feed into the recognition system that generates lexical representations. In addition, to account for L2 development, Escudero and Boersma (2004) and Escudero (2005) extended the LP framework and proposed the Second Language Linguistic Perception (L2LP) Model, which was later revisited and revised by van Leussen & Escudero (2015) (Figure 2.2).

Figure 2.2

The Levels of Representation and Connection Types in the L2LP Model (van Leussen & Escudero, 2015)



Note. This figure, retrieved from van Leussen and Escudero (2015, p. 5), represents an updated visualisation of the original L2LP model proposed by Escudero (2005) with psycholinguistically familiar terminology, as noted by the authors. The 2015 revision further allows for the interaction between perception and recognition, relaxing the earlier assumption of strictly sequential processing.

The L2LP model assumes that perception unfolds across four interconnected levels –

acoustic, phonetic, phonemic, and lexical – each linked by weighted connections[7] (cf. Figure 2.2). For example, in van Leussen and Escudero's (2015) simulation, Dutch learners of Spanish adjusted their mappings from the Dutch three-way contrast /i/-/ɪ/-/ɛ/ to the Spanish two-way /i/-/e/ contrast by weakening connections to the L1-specific intermediate /ɪ/ category over time, thereby achieving more target-like perception. Crucially, this development was not regarded as being prompted by explicit phonological knowledge but instead emerged through lexical (mis)recognition (i.e. learners restructured their perceptual grammar when word recognition failed – a process referred to as meaning-driven learning).

Importantly, the L2LP model makes three foundational claims about the trajectory of L2 perceptual development. First, it posits *Full Copying*, whereby L2 learners transfer their L1 perception grammar to the L2 at the initial stage. Some evidence for this comes from Escudero (2001), who examined Spanish learners of Scottish English (SE) and Southern British English (SBE). It was found that learners of SE (in which /i/ and /ɪ/ have more similar duration but different height) perceived the /i/-/ɪ/ contrast in a more native-like way than learners of SBE (in which /i/ and /ɪ/ have more similar height but different duration). The Spanish learners' reliance on vowel height – a dominant cue in Spanish – indicated that their L2 perception grammar was initially shaped by L1-derived cue-weighting strategies.

Second, the model claims *Full Access* to the same learning mechanisms available in L1

[7]This notion of connection strengths was introduced by van Leussen and Escudero (2015) as a more plausible and flexible mechanism for modelling L2 learning, as it allows connections to be both strengthened and weakened in response to perceptual outcomes.

acquisition, such as 'auditory-guided category formation and lexicon-guided boundary shifting for phonological categories' during L2 development (Escudero, 2007, p. 127). For instance, Spanish learners of SBE were shown to gradually incorporate durational cues – initially novel and non-salient in their L1 – into their perception of the /i/-/ɪ/ contrast with increased experience and proficiency. This shift indicated that learners were able to reweight phonetic cues and refine their perception of L2 categories over time using mechanisms akin to those employed in L1 acquisition (Escudero, 2005).

Third, the model predicts *Full Proficiency*, suggesting learners can attain optimal perceptual categorisation in both L1 and L2 (i.e. demonstrating native-like categories in both languages) at advanced levels. Supporting evidence for this claim was provided by Escudero and Boersma (2002), who examined Dutch learners of Spanish – a case where the Spanish /i/-/e/ contrast is considered a subset of the Dutch vowel /i/-/ɪ/-/ɛ/ contrasts. In a series of perception tasks, learners at more advanced proficiency levels showed reduced reliance on the intermediate Dutch vowel /ɪ/ and demonstrated native-like Spanish two-way distinction patterns. Although the study was cross-sectional, the observed correlation between experience and perceptual alignment appeared to support the model's prediction that native-like perception in the L2 is achievable over time with diminished L1-induced perceptual interference.

Beyond accounting for the mapping of phonetic representations, a notable strength of the L2LP model is its treatment of lexical recognition as the functional endpoint of perception. Unlike many earlier models which focus on pre-lexical phonemic categorisation, it incorporates a pathway from acoustic input to lexical form, addressing how perception may interact with recognition. This multi-layered structure considers not only cross-

language phonetic contrasts but also target language input where learners naturally engage in both segmental and lexical dimensions. Moreover, in its revised version (van Leussen & Escudero, 2015), this model emphasises meaning-driven learning – whereby perceptual category restructuring is driven by communicative success – which guides perceptual learning. In other words, errors at the lexical level prompt updates across four representational levels (cf. Figure 2.2) and gradually refine the learners' perceptual system to align with the L2 norm.

This account is therefore useful for understanding the scope of the current study, where speech perception and word recognition are investigated (cf. Section 2.2.1) with respect to context-dependent features (here CSPs). Despite its breadth and computational advantages, the L2LP model has primarily been applied to segmental contrasts in laboratory settings. While its framework could, in principle, accommodate allophonic variation and context-dependent cue weighting, its current implementation does not directly address the kinds of inter-word phonological variability that characterises CSPs. As such, its capability to account for the perception of dynamic, spontaneous speech – where CSPs play a crucial role – remains partial.

Compensation Hypothesis

Focusing specifically on the perception of phonological variation, Darcy et al. (2007) investigated how L2 learners compensate for assimilation processes. As a CSP category, assimilation refers to context-dependent modification in which certain phonetic properties of a segment shift towards those of an adjacent sound – for example, in English, a word-final /n/ may be realised like /m/ before a bilabial onset, as in *green box* pronounced as /griːmbɒks/ (cf. Section 2.1.2: *Modification* for this CSP category). L1

listeners typically compensate for such alterations unconsciously, mapping the surface variants onto to their intended lexical forms. Darcy et al. proposed that this process of compensation is language-specific and rule-governed, thereby posing a marked challenge for L2 listeners who may not yet possess such phonological knowledge in the target language.

Across a series of word-detection experiments, Darcy et al. (2007) found that beginning learners of both French and English exhibited a higher degree of compensation for the type of assimilation readily present in their L1 compared to those that exist in their L2. That is, native English speakers compensated more for place assimilation in English (e.g. *clean pan* realised as *clea/m/ pan*), while native French speakers compensated more for voice assimilation in French[8] (e.g. *robe sale* realised as *ro/p/ sale*[9]). Notably, these learners continued to adopt their L1 compensation patterns even when listening to the L2 (i.e. L1-English listeners processing L2-French and vice versa), where their corresponding L1 assimilation mechanism did not apply. In contrast, advanced learners demonstrated greater sensitivity to the appropriate assimilation rule in the L2, showing compensation patterns more closely aligned with those of native listeners. Importantly, they did so while maintaining L1-specific compensation in their native language, suggesting L2-specific perceptual compensation rules can be developed without overwriting the native compensation mechanism. Based on these findings, the authors postulated dual processing systems: one responsible for 'parsing continuous speech into phonetic categories', and the other for 'matching phonetic surface forms onto their

[8]The authors noted that for both types of assimilation, the target phonemes undergoing the phonological alterations were shared across English and French.

[9]*Robe* /ʀɔb/ means 'dress' (noun), and *sale* /sal/ means 'dirty' in French.

underlying lexical forms' (pp. 25-26). Notably, the latter was considered sensitive to language-specific and context-dependent phonological rules, such as those governing place or voice assimilation in their stimuli.

Building on this, Darcy et al. (2009) further examined both native and non-native compensation mechanisms in listeners' L1 contexts. Specifically, native French and English speakers were tested on both place and voic assimilation conditions embedded in L1 sentences, with contexts manipulated to be either viable (i.e. phonologically appropriate) or unviable. Results indicated that listeners compensated significantly more for the assimilation type existing in their L1 – i.e. English speakers compensated more for place assimilation (viable condition) than voicing assimilation (unviable condition) when processing English, and vice versa for French speakers. Interestingly, a smaller but significant degree of compensation was observed even for the non-native assimilation type, which may be attributed to language-general auditory mechanisms rather than to phonological inference per se. In light of these findings, the authors proposed a three-stage hypothesis of speech processing: (1) an initial stage where phonetic representations are derived via *language-independent* mechanisms; (2) a subsequent stage where *language-dependent* mechanisms operate to compensate for phonological alternations; and (3) a final stage where the output 'phonological representations are matched against the lexical representations for word recognition' (Darcy et al., 2009, p. 296).

Taken together, this compensation hypothesis offers a more dynamic perspective of speech perception that moves beyond static segmental identification, drawing attention to context-sensitive phonological alterations across word boundaries. It also foregrounds the idea that L2 listening difficulties may arise not only from inaccurate segmental perception

but also from the misapplication, or absence of, language-specific compensation strategies. While empirical support thus far has focused predominantly on the assimilation process, the framework holds broader implications for connected speech perception. In particular, it points to the potential for exploring how learners might be trained to recognise and process the fluidity of connected speech, although further research is needed to extend these insights to other CSP categories and learner populations.

Reduced Forms Perception Model

A study by Wong et al. (2017) is particularly noteworthy for its focused investigation into the cognitive-linguistic underpinnings of L2 learners' perception of connected speech. The authors proposed an integrated, data-driven model that not only illustrates how various phonology-related aspects influence the perception of reduced forms[10] in L2 learners, but also connects these perceptual skills to broader listening comprehension. Specifically, their model centred around two outcome variables – reduced forms perception and general listening comprehension – which were empirically measured through a sentence dictation task[11] featuring reduced forms and a multiple-choice comprehension test based on authentic news and conversation recordings, respectively. Five cognitive-linguistic skills were hypothesised to contribute to these outcomes:

- **Part-word recognition**, as measured by a speech gating task
- **Phonemic awareness**, as measured by tasks involving phoneme deletion, segmentation, blending, and reversal

[10]The term 'reduced forms', as adopted in Wong et al. (2017), is an alternative label for 'connected speech processes (CSPs)', and it is used here when referring to their model.

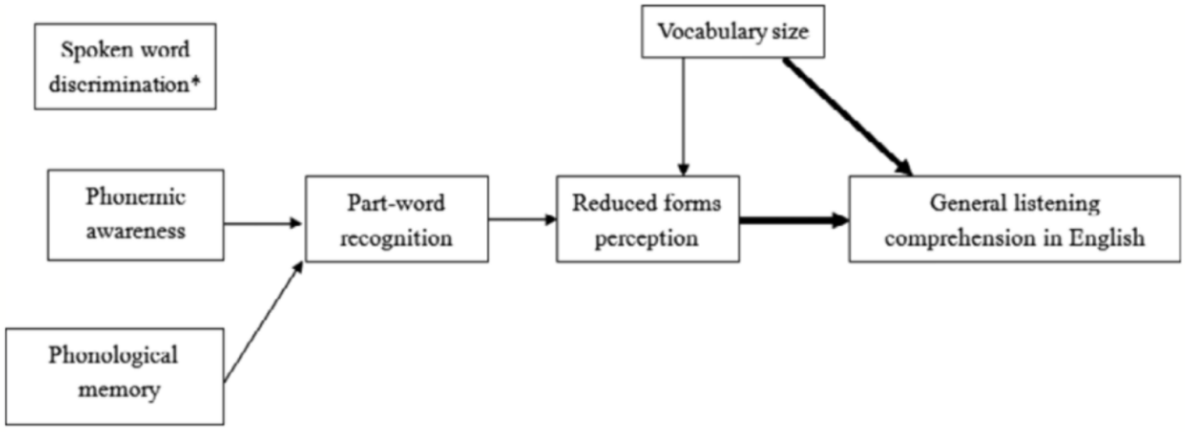
[11]The dictation task comprised 33 sentences representing nine types of reduced forms: contraction, juncture, elision, vowel weakening, assimilation, intrusion, flapping, glottalisation, and palatalisation.

- **Phonological memory**, as measured by a non-word repetition task
- **Spoken word discrimination**, as measured by a minimal pair discrimination task
- **Receptive vocabulary**, as measured by a synonym identification task

Using hierarchical regression analyses, Wong et al. found that both reduced forms perception and receptive vocabulary significantly predicted L2 learners' general listening comprehension. Additionally, part-word recognition and vocabulary size emerged as direct predictors of reduced forms perception, underscoring their critical roles in decoding dynamic phonological alterations in connected speech. In contrast, phonemic awareness and phonological memory exerted indirect effects on reduced forms perception, mediated by part-word recognition, suggesting a developmental hierarchy among these perceptual skills. The authors provided a schematic summary of their analytical results as presented in Figure 2.3, retrieved from Wong et al. (2017, p. 24).

Figure 2.3

The Schematic Representation of Reduced Forms Perception and Listening Comprehension (Wong et al., 2017)



Note. This figure, retrieved from Wong et al. (2017, p. 24), represents their statistical

analyses of the variables. Bold arrows indicate predicting effects on general listening comprehension; non-bold arrows indicate direct or indirect predicting effects on reduced forms perception. An asterisk '*' denotes an absence of significant regression path.

Crucially, this model moves beyond studies focused on one specific dimension of perception by incorporating a spectrum of linguistic and cognitive metrics – from acoustic-phonetic (e.g. part-word recognition and phonemic awareness) to phonological processing (e.g. phonological memory and reduced forms perception), and ultimately to listening comprehension levels. It thereby reflects the layered demands of real-time speech processing. More importantly, the model explicitly addresses reduced forms (i.e. CSPs), serving as a theoretical foundation for exploring the perceptual challenges that L2 listeners encounter in understanding naturally and spontaneously occurring speech.

While the model provides an informative framework for understanding reduced forms perception, it does not distinguish between the impact of different types of reduced forms, likely due to the limited number of stimuli (33 items distributed across nine types of reduced forms). Nonetheless, it marks an advance by bridging theoretical constructs of speech perception with the practical complexities encountered by L2 listeners in connected speech contexts. One key pedagogical implication is that multiple layers of processing – rather than merely segmental accuracy – must be considered to support L2 learners' perception of connected speech and broader listening comprehension.

In sum, throughout the evolution of speech perception models and hypotheses, an increasing amount of research has shifted from modelling isolated phoneme acquisition to exploring how cognitive-linguistic mechanisms operate at different stages of speech

processing, including the perception of phonological variation in connected speech. Models such as the SLM-(r) and L2LP-(r) have provided foundational insights into segmental development, yet offer comparatively limited accounts of the dynamic, context-dependent nature of connected speech. On the other hand, the phonological compensation hypothesis (Darcy et al., 2007, 2009) and the reduced forms perception model (Wong et al., 2017) have begun to reveal how different phonological correlates influence L2 listeners' perception of CSPs and broader listening comprehension. Building upon these developments, the current study investigates the phonological interactions between individual segments across word boundaries in learners' L2 speech perception. It also seeks to generate richer empirical data across a wider range of CSP categories, with the aim of probing context-dependent mechanisms of phonological processing (e.g. Darcy et al., 2007; 2009) and contributing further insights into the theoretical base underpinning the perception of L2 connected speech (e.g. Wong et al., 2017). The methodological framework employed to pursue these objectives is outlined in Chapter 3:.

2.2.3 Top-Down and Bottom-Up Processing

Alongside the development of models and hypotheses in speech perception, the distinction between top-down and bottom-up processing strategies has also been an important concept that influences language pedagogy. This section reviews the essence of these two approaches, focusing on their implications for L2 speech perception, as a prelude to discussing teaching and learning practices.

In comprehending spoken language, there has been an ongoing debate regarding the relative importance of top-down and bottom-up processing. Top-down processing, by and large, refers to leveraging contextual resources such as prior knowledge, expectations,

and familiarity with language patterns to predict or infer the speaker's meaning (Field, 1999). This approach allows listeners to use their broader understanding of the communicative situation to interpret a speech stream, essentially drawing on hints from a higher level of linguistic representation (i.e. the global meaning at the 'top') to anticipate what is likely to be said[12].

In contrast, bottom-up processing involves real-time linear decoding of the speech signal (Lynch, 2006). This approach, typically considered data-driven (Field, 1999), depends on listeners' ability to perceive and integrate acoustic information and linguistic knowledge as the speech unfolds. That is, listeners engaging in bottom-up processing sequentially piece together various levels of signal – starting from phonemes, morphemes, words, building up to phrases, clauses, sentences, and discourses. Throughout this process, listeners arrive at the overall meaning of the utterance incrementally upwards from the lower-level linguistic elements at the 'bottom'.

It is worth noting that substantial evidence indicates that top-down processing is particularly important for L1 listeners, who typically anticipate acoustic input and resolve ambiguities by drawing extensively on contextual information (e.g. Warren, 1970; Samuel, 1981; Norris et al., 2003). Historically, likely influenced by relatively robust L1 evidence, L2 pedagogy during the 1980s and 1990s privileged top-down processing (Field, 2008a). This instructional preference also aligned chronologically with schema theory (Rumelhart & Ortony, 1977; Rumelhart, 1980) and communicative language

[12]It is acknowledged that, broadly speaking, any use of information from a higher-level unit (e.g. lexical items) to facilitate the recognition or identification of a lower-level unit (e.g. individual phonemes) may also be considered top-down processing.

teaching (e.g. Canale & Swain, 1980; Canale, 1983), both of which are consistent with an emphasis on inference-making and global understanding. Practitioners tended to assume that when learners encountered unfamiliar spoken texts, they could effectively rely on contextual cues to overcome difficulties in decoding linguistic details in the input. As Flowerdew and Miller (2005) noted, learners were believed to be 'quite able to identify truncated words so long as they are presented with the surrounding context' (p. 25), and their prior experience, including their understanding of the world, knowledge of discourse structures, and familiarity with pragmatic contexts, 'compensates for any problems in understanding microlevel elements, such as sound discrimination, syntax, and word and utterance level semantics' (p. 26).

However, empirical research has increasingly questioned the sufficiency of relying predominantly on the top-down approach for L2 learners. A pivotal large-scale study by Tsui and Fullilove (1998) examined Hong Kong ESL learners' listening performance across two types of schema-based materials. In the 'matching' schema type, the initial cues activated a mental representation (schema) that remained consistent with the subsequent input, thus allowing listeners to rely on top-down processing to infer the correct response. In contrast, the 'non-matching' schema type introduced an initial schema that was later contradicted deliberately by incoming information, thereby placing heavier demands on bottom-up processing for accurate decoding of continuous speech input. To determine which processing strategy better distinguished more skilled from less skilled listeners, the authors analysed *mean criterion scores* – a metric reflecting the average total score of candidates who selected each response option. A higher mean criterion score for the correct response indicated that more proficient listeners tended to choose that option. The authors found that 'non-matching' schema items consistently

elicited higher mean criterion scores than 'matching' schema items, indicating that those who accurately responded to the 'non-matching' items demonstrated higher overall listening proficiency. Consequently, they concluded that bottom-up processing, rather than top-down inference, played a more critical role in differentiating more skilled from less skilled L2 listeners.

The growing recognition of the limitations of an overreliance on top-down processing has prompted, in comparatively recent times, a resurgence of interest in bottom-up processing among practitioners and researchers (Field, 2008a; Flowerdew & Miller, 2005). Through closer examination of word recognition mechanisms, academics have urged that top-down processing alone is insufficient without the assistance of bottom-up decoding abilities (e.g. Magnuson, 2017; Vandergrift, 2004). Field (2003) specifically highlighted that one of the most pressing challenges for L2 listeners is the accurate demarcation of word boundaries, a task that is particularly difficult in spontaneous connected speech. He proposed that exercises such as dictating ambiguous utterances can raise students' awareness of phonological variability, and in turn determine the levels where errors occur. This emphasis upholds the promotion of bottom-up processing given that 'once a word is spoken next to other words, the way it is pronounced is subject to a wide variety of processes' (Alameen & Levis, 2015, p. 2).

Indeed, one domain where bottom-up processing appears to play a particularly salient role is the perception of connected speech processes (CSPs). These phonological processes (cf. Section 2.1.2 for an overview of various CSP categories) tend to obscure the citation forms of words when they are spoken in naturalistic discourse. As a result, the surface forms (containing CSPs) listeners encounter may diverge considerably from

their expected citation forms, thereby disrupting processing fluency. It is therefore an empirical question how, and to what extent, different types of training or input exposure may help L2 listeners develop sufficient bottom-up processing skills to deal with speech characterised by extensive CSPs. However, relatively few studies have specifically emphasised the importance of bottom-up processing in addressing CSPs using tightly controlled measures, nor have different training approaches been systematically evaluated regarding their effectiveness in helping learners accurately perceive speech affected by particularly challenging CSPs. Importantly, while CSPs may point to the need for robust bottom-up skills, this does not imply that top-down processing is irrelevant. Contrary to traditional dichotomous frameworks that view bottom-up and top-down processing as distinct or opposing approaches, contemporary perspectives emphasise their dynamic and interactive nature (e.g. Buck, 2001; Siegel, 2018) – i.e. neither is sufficient on its own. In practice, listeners predict meaning based on contextual cues while simultaneously updating these predictions with incoming speech input. Thus, the bottom-up and top-down approaches are not mutually exclusive, but rather functionally interdependent components of speech perception.

In the context of the current study, which investigates how L2 learners perceive utterances affected by CSPs, we recognise that both processing routes inevitably interact. However, given the specific aim to pinpoint perceptual difficulties induced by CSPs, our design places particular emphasis on learners' bottom-up decoding abilities, in responses to previous calls for more bottom-up training in L2 listening instruction (e.g. Field, 2003; Hulstijn, 2003; Wilson, 2003). Specifically in the current study, a principal measure – a tightly controlled dictation task – deliberately minimises contextual predictability (the linguistic measure and stimulus design are detailed in Sections 3.3.1 and 4.1.3,

respectively), thereby providing clearer insights into listeners' phonological decoding abilities. While this methodological focus underscores bottom-up processing, again we do not claim that top-down mechanisms are absent. In line with current theoretical perspectives, we acknowledge that both processing approaches are likely engaged, and hence address potential top-down influences in the interpretation of (mis)perception patterns observed in the study (see Section 6.1).

2.3 Teaching and Learning Practices

This section steers the discussion towards pedagogical practice by reviewing first the listening aspect of language teaching and learning, and second instruction in CSPs, followed by comments on the methodological challenges in CSP studies.

2.3.1 The Skill of Listening

Before discussing the empirical evidence for instruction in CSPs, this section provides a focused review of the role of listening in L2 teaching and learning, along with the evolving pedagogical priorities in this domain.

As noted by Vandergrift (2004), listening is probably the most challenging language skill (for L2 learners) to master. For many learners, particularly those at lower proficiency levels, listening is often perceived as the area in which they feel least successful (e.g. Renandya & Farrell, 2011). This difficulty is commonly associated with the inherently ephemeral nature of listening comprehension, wherein the cognitive processes involved are not directly observable, in contrast to other skills such as reading and writing, which are anchored in visual input or output and can benefit from the referential stability of written texts. This struggle that learners experience in acquiring listening skills is often

mirrored by teachers, who may likewise encounter difficulties in designing and implementing listening instruction (Walker, 2014; Martínez-Flor & Usó-Juan, 2006). Further compounding this issue is the fact that 'since these processes are covert, listening is a difficult skill to *research*' (Vandergrift, 2004, p. 18, emphasis added). As a result, in comparison to other language skills, listening remains a relatively understudied aspect of second language pedagogy.

Due to these challenges, listening has long been under-emphasised in language classrooms and often deemed a passive language skill or activity that requires little pedagogical attention (Osada, 2004). In many educational contexts, it has been observed that listening is frequently *tested* but not actually *taught* (e.g. Sheerin, 1987; Mendelsohn, 2006). This long-standing neglect stands in stark contrast to the reality that listening is likely the most frequently used skill, accounting for the largest proportion of language use in everyday life (e.g. Burley-Allen, 1995; Rankin, 1928; Rost, 2001). Over the years, this discrepancy between the real-life significance of listening and its peripheral role in instruction has prompted growing calls for more deliberately teaching the *process* of listening rather than merely testing the *product* of listening (e.g. Vandergrift, 2004; Field, 2008b).

In the domain of pronunciation instruction – encompassing both speaking and listening – segmental features (i.e. vowels and consonants) have traditionally been prioritised as the fundamental unit to start from (Wang, 2022). Therefore, the perception of cross-linguistic segments has been the focus of extensive empirical inquiry (e.g. Brown, 1998; Flege & Wang, 1989; Michaels, 1974). On the other hand, while segmental precision remains an important component, another body of research has highlighted the pivotal role of

suprasegmental features (e.g. Anderson-Hsieh et al., 1992; Derwing & Rossiter, 2003) – including lexical stress (e.g. Cutler, 2015), intonation (e.g. Grabe et al., 2003), and rhythm (e.g. Gilbert, 2019). These prosodic components are found to be particularly crucial in parsing speech streams and interpreting speaker intent, and as such should not be ignored in successful listening instruction.

Beyond the conventional distinction between segmental and suprasegmental features, CSPs represent a related but underexplored aspect of listening pedagogy – as they involve interactions of segmental and suprasegmental features (Alameen & Levis, 2015) (cf. Section 2.1.2 for a review of CSP categories and their phonological mechanisms). Early studies have indicated that such phonological alterations can substantially hinder L2 learners' speech perception (e.g. Brown & Hilferty, 1986; Henrichsen, 1984; see Section 2.3.2 below for a review of empirical findings on CSPs). Nevertheless, many instructional materials and textbooks have overlooked CSPs or presented them only sparingly, focusing instead on enunciated, 'standard' speech forms that deviate from the spontaneous pronunciation patterns in authentic discourse.

More recently, however, there has been increasing advocacy for the inclusion of CSPs in pronunciation and listening instruction, often calling for, at the very least, raising learners' awareness of CSPs as they occur in everyday spoken language (Ahmadian & Matour, 2014). Cahill (2006) offered a critical perspective on this issue, arguing that 'failure to address features that are masked by discourse adjustments is a dereliction of duty by the teacher, and it will eventually leave students with serious communicative deficiencies' (p. 99). Reflecting this trend, CSPs have been formally incorporated as a core component within the pronunciation instruction model proposed by Reed and Michaud (2011), which

underscores their influences on L2 listening development. Despite the growing recognition of the importance of CSPs, their pedagogical treatment still requires more empirical research to better inform teaching practices and guide future investigations (Momen & Pilus, 2022).

Overall, as research continues to explore in greater depth how L2 learners develop listening proficiency, CSPs have gained increasing attention (beyond the conventional distinction between segmental and suprasegmental features) for both their theoretical implications and potential instructional utility. The following section reviews empirical studies that investigate L2 learners' perception of connected speech, with a particular focus on the effectiveness of CSP-focused instruction.

2.3.2 Instruction in CSPs

An early study on CSPs by Henrichsen (1984) used a sentence dictation task to assess the impact of Sandhi-variation[13] on spoken word recognition among high-level and low-level English learners primarily with Asian-Pacific backgrounds (e.g. Hong Kong, Japan, Korea, the Philippines, and Samoa). The results indicated that the performance of both learner groups significantly improved in the absence of Sandhi-variation, such as contraction, reduction, or assimilation, while native English listeners were unaffected. Expanding upon Henrichsen's work, Ito (2001) distinguished between (a) lexical reduced forms[13] which involve some alteration in pronunciation when contracted with the negating adverb *not* (e.g. *doesn't* and *won't*), and (b) phonological reduced forms which

[13]The terms 'Sandhi-variation' and 'reduced forms' are used as alternatives to 'connected speech processes (CSPs)'. References to those terms reflect the original terminology used in the respective sources, including those reviewed throughout this section.

involve omission of sounds in contracted copulas or auxiliaries (e.g. *she's* and *I've*). Ito reported similar findings that both intermediate and advanced learners (speaking different Asian languages as their L1, including Japanese, Korean, and Chinese) performed better in sentence dictation when reduced forms were absent, with native speakers showing no significant difference. Notably, learners were found to struggle more with phonological reduced forms compared to lexical reduced forms.

While these two studies highlighted the challenges L2 learners face in perceiving utterances containing CSPs, they also revealed methodological constraints. Their stimuli were designed with target phrases placed in fixed positions within similar sentence structures, potentially underestimating the difficulty learners may encounter during authentic communication where CSPs manifest in varied pragmatic and syntactic contexts. Additionally, the CSP categories were narrowly defined and selected, thereby neglecting many frequently used CSPs not represented in their written forms. Importantly, these were not intervention studies that could directly inform pedagogical practice or learners' developmental trajectories.

In addressing pedagogical interventions, Brown and Hilferty (1986) assessed the effects of short daily lessons (less than ten minutes each) focused on reduced forms for Chinese intermediate EFL students through measuring their performance on the Integrative Grammar Test [14] (IGT; Bowen, 1976), reduced forms dictation, and listening

[14]The IGT consists of sentences that contain a target word (typically the second word) with reduced clarity 'through normal reductions, assimilations and contractions of informal spoken English' (Bowen, 1976, p. 30). The examinees are asked to identify and write down the citation form of the target word after listening to the sentence based on the lexical-syntactic context. For example, the intended answer for *Who'd he been to see?* is the auxiliary *had* (instead of *did* or *would*). Likewise, in *Give 'm an inch and he'll take a mile*,

comprehension tasks. An experimental group (n = 16) was provided with instruction in reduced forms, and a control group (n = 16) received placebo instruction in other pronunciation features such as stress and intonation. It was found that the experimental group outperformed the control group in the IGT and dictation tasks, but not in listening comprehension. However, the content of instruction was not strictly controlled, and caution was advised by the authors against potential practice effects in that the same type of dictation was used during both the instruction and the assessment for the experimental group. Further, because the details of stimuli are not available in the publication[15], it is unclear which CSP categories were investigated and how many items were included in each category.

Targeting a wider range of reduced forms, Matsuzawa (2006) examined ten categories including flapping, linking (C-C, C-V, V-V), deletion, assimilation, weak forms (of function words), contraction, glottalisation, /nt/ reduction, and high-frequency phrases (e.g. *wanna* and *gonna*). Eight 30-minute lessons were designed for a beginner and an intermediate group of Japanese adult EFL learners (N = 20). Significant improvement in sentence transcription from a pre-test to a post-test was reported, and interestingly the beginner group benefited equally from the lessons as the intermediate group. Despite greater inclusiveness as to the range of reduced forms investigated, the lack of a control group and limited stimulus items (3 items per category) were notable limitations.

Using a dictation cloze test for assessment, Carreira (2008) focused on instruction in

---

the intended answer is the singular pronoun *him* (instead of *them*).

[15]Although the authors referenced an appendix containing stimulus information, it was not included in the article retrieved from SAGE Publications.

contraction, palatalisation, and vowel reduction for adult EFL learners in Japan with diverse Asian backgrounds (N = 19). The instruction included explanation of reduced forms and dictation exercises built from conversations and pop songs. The results showed significant improvement in participants' word recognition (as measured by a fill-in-the-blank dictation task) but not in their general listening ability (as measured by TOEIC listening sections). In this study, again, there was no control group. Moreover, while the author reported that the course comprised thirty weeks of instruction in reduced forms, neither the course structure nor detailed instructional content was provided.

Adopting a similar measure of cloze dictation (also referred to as fill-in-the-blank task, with variations of the term used across different studies), Ahmadian and Matour (2014) and Rahimi and Chalak (2017) observed improved dictation performance among Iranian EFL learners following explicit instruction in reduced forms, compared to control groups. Ahmadian and Matour targeted intermediate adult learners (n = 20 for each group) and focused on six categories of CSPs (weak forms, place assimilation, voice assimilation, palatalisation, elision, intrusive *r*), while Rahimi and Chalak targeted pre-intermediate learners aged 14 – 19 years (n = 25 for the experimental group, n = 27 for the control group) and did not specify the CSP categories investigated. In contrast to Brown and Hilferty (1986), the experimental group in Rahimi and Chalak (2017) attained significantly higher scores than the control group not only in word recognition but also in listening comprehension after the pedagogical treatment. Although both Ahmadian and Matour (2014) and Rahimi and Chalak (2017) reported employing an 'explicit' instructional approach for their experimental groups, the nature of the instruction was not elaborated in detail – only generic descriptions were provided regarding the teaching procedures, including mentions of 'various types of instruction' or the use of instructional

tools such as a laptop, a handout, and clips featuring native speakers.

Exploration of different instructional approaches to teaching CSPs can be seen in a few studies. For instance, Abe (2010) compared Negotiation of Form (NoF; Lyster & Ranta, 1997, as cited Abe, 2010) and rule explanation in Japanese college EFL classes for beginner to intermediate learners. NoF aimed to encourage learners to contrast speech produced with and without CSPs and discuss what they noticed, as opposed to providing learners with rules to follow. It was found that the NoF group exhibited greater improvement in all four investigated features (rhythm, linking, assimilation, elision) than the rule explanation group. Replicating Abe's study, Kul (2016) examined similar instructional contrasts – labelled as inductive (NoF) versus deductive (rule explanation) methods – with advanced Polish EFL learners, focusing on vowel reduction, palatalisation, /t/ elision, and /h/ elision. In contrast to Abe, Kul found no significant differences between the two methods in enhancing learners' perception of CSPs, despite both groups' overall improvement. Several factors may be speculated to have caused the different results between the two studies, such as learners' proficiency levels, L1 backgrounds, the CSPs investigated, and the assessment measures[16].

Particularly noteworthy are studies by Ting and Kuo (2012) and Kuo et al. (2016), which focused on learners with a similar language background as in the current research. Ting and Kuo targeted Taiwanese sophomore English majors and examined six CSP categories (C-V linking, elision, palatalisation, contraction, /h/-deletion, and flapping). They found

[16]The only assessment information reported in Abe (2010) is that 'the pre- and post-test consisted of 20 questions, including the targeted prosodic features' (p. 2). No further details are provided regarding the test procedures or question types. In contrast, Kul (2016) employed a sentence dictation task to assess learners' perception of words affected by CSPs.

that an experimental group which received explicit instruction supported by pop song lyrics outperformed a control group (with no instruction) in a sentence dictation task at a post-test. Kuo et al. targeted Taiwanese junior high school students and compared the effects of explicit and communicative instruction in linking, elision, palatalisation and contraction. The explicit group was taught the rules of CSPs with the aid of phonetic symbols, while the communicative group engaged in guided learning tasks (e.g. creating and performing dialogues in pairs) in which CSPs were embedded. Similar to Kul (2016), no significant difference was found between the two approaches in improving learners' ability to recognise words in connected speech, although in Kuo et al.'s study both experimental groups surpassed a control group receiving no instruction in CSPs.

Overall, some literature has suggested promising effects of explicit CSP instruction compared to no instruction (e.g. Brown & Hilferty, 1986; Matsuzawa, 2006; Ahmadian & Matour, 2014; Rahimi & Chalak, 2017) on enhancing learners' word recognition in connected speech. However, other studies have reported evidence in support of more implicit or communicative instruction (e.g. Abe, 2010; Kuo et al., 2016). Consequently, the relative effectiveness of different pedagogical approaches, particularly for specific learner populations, remains an area of ongoing debate and in need of further research.

2.3.3 Methodological Challenges in CSP Studies

Despite ongoing efforts to accumulate evidence for instruction in CSPs, several prevalent methodological issues have emerged alongside inconsistent findings from the prior studies. These warrant a discussion and underscore the necessity of further rigorous investigation, as outlined below.

Absence of Diagnostic Procedures

A prominent methodological issue is the absence of diagnostic procedures prior to interventions. Studies tend to select certain CSPs as targets for interventions without first empirically verifying whether these categories genuinely impose perceptual difficulties on the intended learner population. Therefore, the gravity of the difficulties is often not assessed, thereby limiting the capacity to design interventions tailored to learner needs. Without a systematic diagnosis, the selection of CSPs in earlier studies may be speculatively based on personal experience, research interest, or merely underrepresentation in the existing literature, rather than on evidence-driven criteria.

Limited Scope of CSP Categories and Stimulus Inventory

A related issue pertains to the range of CSP categories and the number of stimuli included for investigation. Most intervention studies have targeted fewer than five CSP categories, with exceptions being Matsuzawa (2006), Ting and Kuo (2012), and Ahmadian and Matour (2016). While restricting the scope to fewer CSP categories may allow for less demanding intervention design, including a broader array of CSPs enables comparisons of how various categories affect learners' speech perception to different extents. Further, studies utilising only a small number of stimulus items per CSP may have limited power to evaluate category-specific effects. For example, Matsuzawa (ibid.), despite covering ten CSP categories, included only 3 stimulus items per category, thus limiting the strength and generalisability of the findings related to individual categories.

Insufficient Control Over Intervention Input

Additionally, stringent control over instructional input has been frequently overlooked, in terms of both content and sequencing. In L2 learning contexts, any linguistic or

instructional elements – ranging from target words and sentences to presentation formats and instructional wording – may function as input. Consequently, these elements should be meticulously controlled across intervention conditions to ensure precise evaluation of pedagogical efficacy and mitigate confounding influences arising from unequal or unintended exposure. Further, the sequencing of instructional content is rarely balanced or reported, introducing potential biases, such as recency effects (wherein learners tend to retain recently presented materials), particularly in post-intervention assessments. Unfortunately, these intervention details have often been insufficiently documented in previous studies (e.g. Carreira, 2010; Abe, 2010), and no existing CSP studies (to my knowledge) have explicitly addressed the issue of instructional sequencing.

Lack of Differentiation Between Intelligibility and Comprehensibility

Finally, although many intervention studies employ intelligibility measures (typically operationalised as participants' ability to recognise CSP-affected words in connected speech) and listening comprehension more broadly (e.g. Brown & Hilferty, 1986; Carreira, 2008; Kuo et al., 2016; Rahimi & Chalak, 2017), few have differentiated between the constructs of intelligibility and comprehensibility as outlined by Munro and Derwing (1997). Intelligibility and comprehensibility, though moderately correlated, capture distinct dimensions of speech perception. The former concerns the accuracy with which a listener identifies intended words within an utterance, while the latter relates to the level of ease or difficulty a listener experiences in understanding an utterance (see Section 3.2 for a detailed discussion). This conceptual distinction is particularly relevant for connected speech research, where both accuracy and perceptual effort may be independently affected. Therefore, incorporating these constructs separately in evaluating the effectiveness of CSP instruction may offer more nuanced pedagogical insights.

To address these methodological issues, this research (i) incorporates a diagnostic phase to identify CSP categories that are particularly challenging for the target learner population, and inform the design of subsequent pedagogical interventions; (ii) devises a more extensive and representative stimulus inventory encompassing a broader range of CSP categories; (iii) implements meticulously controlled instructional materials to enable more reliable comparisons of the effectiveness of different intervention conditions; (iv) employs separate linguistic measures for intelligibility, comprehensibility, and general listening comprehension to capture distinct dimensions of speech perception among L2 learners. These strategies, intended to enhance the methodological rigour and validity, will be further elaborated in Chapter 3:.

2.4 Summary of Literature Review

In summary, when utterances are produced in naturalistic contexts, various phonological alterations – termed connected speech processes (CSPs) – systematically change the pronunciations of words compared to their citation forms. These alterations often pose challenges to L2 learners' speech perception. Speech perception, in this regard, involves a set of cognitive processes that transform auditory input into linguistic representations from which words can be recognised. While a range of theoretical models and hypotheses have been developed to explain how L1 and L2 sounds may be categorised, represented, and identified, few have explicitly addressed the impact of CSPs within L2 listening contexts.

In L2 speech processing, the top-down approach historically held primacy in both research and pedagogy. However, more recent scholarship has increasingly

acknowledged the limitations of relying predominantly on the top-down approach, especially in contexts where listeners encounter dynamic phonological alterations inherent in connected speech. As a result, greater emphasis has appeared to shift towards the bottom-up approach, particularly focusing on the development of learners' phonological decoding skills for processing authentic speech input, including CSP-affected lexical items.

In pedagogical practice, however, listening has long been an overlooked skill in language classrooms, which is typically *tested* but rarely *taught* to L2 learners. Furthermore, CSPs continue to be treated as a peripheral component in pronunciation instruction, compared to conventional segmental and suprasegmental features. Nevertheless, a growing body of research highlights the potential benefits of targeted CSP instruction in enhancing L2 learners' ability to process connected speech. Yet even so, empirical evidence in this area remains limited in both breadth and depth – likely due to various methodological challenges – such that few studies have been able to report the effects of individual CSP categories, and the relative effectiveness of different pedagogical approaches remains largely inconclusive.

Guided by the literature in the field, this research aims to advance the understanding of how CSPs affect L2 speech perception and to explore the effectiveness of different pedagogical approaches, particularly for adult Mandarin-speaking learners of L2 English. More broadly, it also hopes to provide empirical evidence that may inform the development of theories and models of L2 speech perception in consideration of CSPs.

2.5 Research Questions

Following the aims noted above, the specific research questions driving this inquiry are outlined as follows, with RQ1 corresponding to Phase 1 (the diagnostic study) and RQ2 to Phase 2 (the intervention study):

RQ 1. How do English CSPs affect adult Mandarin-speaking learners' L2 speech perception?

RQ 1a. Does the L2 learners' ability to recognise CSP-affected words vary across different categories of CSPs?

RQ 1b. If so, which categories of CSPs are particularly challenging for the L2 learners?

RQ 2. What is the effectiveness of pedagogical interventions targeting CSPs on adult Mandarin-speaking learners' L2 speech perception?

RQ 2a. Do rule-based, noticing-based, and implicit approaches enhance the intelligibility (i.e. transcription accuracy) of connected speech for the L2 learners?

RQ 2b. Do rule-based, noticing-based, and implicit approaches enhance the comprehensibility (i.e. perceived ease of understanding) of connected speech for the L2 learners?

RQ 2c. Following RQ2a and RQ2b respectively, if so, is one approach more effective relative to the others?

RQ 2d. Do the interventions targeting CSPs have a broader effect on the L2 learners' general listening comprehension?

RQ 2e. How do intelligibility, comprehensibility, and general listening comprehension measures correlate with each other before and after the interventions?

Chapter 3: General Methodology

This chapter presents an overview of the research design, introduces two core methodological constructs (intelligibility and comprehensibility), and delineates the linguistic measures employed throughout this inquiry. Detailed methods for the two phases of study will be elaborated in Section 4.1 and Section 5.1, respectively.

3.1 Research Design

Employing an experimental and quantitative paradigm, this research was structured into two distinct but cumulative phases: diagnosis (Phase 1) and intervention (Phase 2). Overarchingly, this research aimed to identify challenging CSP categories and explore potentially effective pedagogical approaches for L2 learners. Both phases targeted adult Mandarin-speaking learners of L2 English, and each corresponded to a research question posited in Section 2.5. Table 3.1 summarises the objectives, main questions, and components of the two phases.

Table 3.1

The Essence of Phase 1 and Phase 2 Studies

	Phase 1 – Diagnosis	**Phase 2 – Intervention**
Objective	To identify challenging English CSP categories for L2 learners	To explore effective pedagogical approaches for teaching CSPs
Main Question	How do different categories of English CSPs affect L2 learners' speech perception?	What is the effectiveness of pedagogical interventions on L2 learners' perception of CSPs?
Components	• Online experiment comprising	• Pre-test: online experiment

	various linguistic measures	• Intervention: distinct online pedagogical training programmes • Post-test: online experiment

Phase 1 focused on examining the impact of CSPs and identifying CSP categories that were particularly challenging for the target learner population, while also examining the correlations between L2 speech perception, overall L2 proficiency, and general L2 listening comprehension. This entailed scrutinising participants' performance in an online experiment comprising various linguistic measures featuring meticulously designed stimuli.

Phase 2 evaluated the effectiveness of pedagogical interventions targeting those CSP categories identified as particularly challenging in Phase 1. This phase involved sequentially administering a pre-test, highly controlled intervention programmes, and a post-test to participants. Both the pre-test and post-test experiments comprised similar linguistic measures as in Phase 1, while during the intervention, participants engaged in CSP training programmes via distinct pedagogical approaches. General and relative effectiveness of each pedagogical approach was to be determined by comparing different groups of participants' gains from the pre-test to post-test on various linguistic measures. The linguistic measures employed across Phase 1 and Phase 2 studies are described in detail in Section 3.3 below.

3.2 Intelligibility and Comprehensibility

Central to this investigation into L2 speech perception are the constructs of intelligibility and comprehensibility. These two constructs have been argued as 'the gold standard for

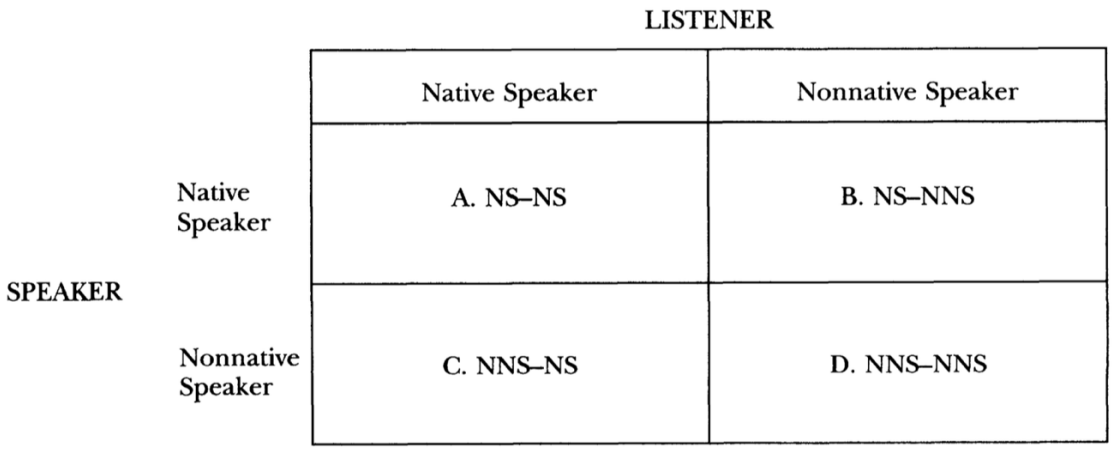
measuring pronunciation improvement' (Derwing & Munro, 2015, p. 90). The current study adhered to the definitions proposed in the seminal studies by Munro and Derwing, who describe intelligibility as 'the extent to which a speaker's message is actually understood' (Munro & Derwing, 1995, p. 76) and comprehensibility as 'how difficult or easy an utterance is to understand' (Munro & Derwing, 1997, p. 2). In other words, intelligibility hinges on the proportion of lexical items (the principal carriers of semantic information) that can be accurately recognised by listeners. In practice, intelligibility is typically measured through relatively more objective metrics, such as transcription accuracy. Comprehensibility, on the other hand, involves a more subjective evaluation of listeners' cognitive effort during speech processing (while how much is actually understood may vary). Munro and Derwing characterise the construct of comprehensibility as the 'perception of intelligibility' or 'perceived comprehensibility' (1997, p. 2; 1995, p. 77), which is typically measured through rating scales on which listeners self-report the level of ease or difficulty in understanding spoken materials.

Although these two constructs may appear closely related, intelligibility and comprehensibility represent distinct dimensions of speech perception. They differ both conceptually (message understood vs. ease of understanding) and methodologically (objective accuracy vs. subjective perception) as noted above. In addition, as the intervention results will show later in this thesis, these two constructs may even exhibit divergent developmental patterns, underscoring the importance of carefully distinguishing them pedagogically (see relevant results in Sections 5.2.2 and 5.2.4, and further discussion in Section 6.3). Levis's (2005) intelligibility principle suggests that different linguistic features may impose varying influences on understanding, and importantly, outlines four possible speaker-listener combinations and their pedagogical

considerations (Figure 3.1). Situated within this framework, the current research specifically examined Quadrant B – i.e. where *non-native (L2) listeners* are assessed for their ability to process connected speech produced by *native (L1) speakers* – with a focus on the impact of CSPs.

Figure 3.1

The Speaker-Listener Intelligibility Matrix (Levis, 2005)



Note. This figure was retrieved from Levis (2005, p. 372).

On the other hand, comprehensibility, sometimes viewed as a measure of processing fluency, reflects a metacognitive dimension of perception (Trofimovich et al., 2024) in contrast to more criterion- and accuracy-based intelligibility. Studies have shown that comprehensibility ratings can also be influenced by various factors, including speaking tasks (e.g. Crowther et al., 2018), listener experiences (e.g. Saito et al., 2019), and social-affective dynamics between speakers and listeners (e.g. Taylor Reid et al., 2019). In general, comprehensibility ratings are found to be relatively consistent among L1 listeners' evaluations of L2 speech (e.g. Kennedy & Trofimovich, 2008) and also between L1 and L2 listeners' evaluations of L2 speech (e.g. Crowther et al., 2016), whereas greater variability has been reported among different L2 listener groups, particularly depending

on whether they share an L1 background with L2 speakers (e.g. Foote & Trofimovich, 2018; Ludwig & Mora, 2017; Mora, 2022). In addition, L2 learners' self-assessments of comprehensibility often diverge from those provided by external L1 listeners; specifically, learners who receive lower ratings from L1 listeners tend to overestimate their speech performance, whereas learners who receive higher ratings from L1 listeners tend to underestimate it (Ortega et al., 2022; Trofimovich et al., 2016). Notably, however, research examining L2 listeners' evaluations of L1 speakers' speech appears comparatively limited, highlighting a gap which this study aimed to address. Furthermore, meta-analytic results from Saito (2021), which targeted evaluations of L2 speakers' speech, identified significant associations between comprehensibility and segmental, prosodic, and fluency features. Given the focus of the current study on CSPs, which inherently influence both segmental and prosodic features, findings from this investigation may provide further empirical insights to this body of research.

To capture these two distinct dimensions of speech perception, the current study included these two constructs as separate dependent variables measured independently. It is worth noting that historically in many prior studies, both intelligibility and comprehensibility have been applied to assess speech produced by L2 learners based on L1 listeners' responses (i.e. Quadrant C in Figure 3.1). In other words, *native (L1) listeners* do the transcription task and rate how comprehensible and accented the speech produced by *non-native (L2) speakers* sounds to the native ears. In contrast, the current study reversed the typical approach by examining the intelligibility and comprehensibility of *L1 speakers'* speech from the perspective of *L2 listeners*, to specifically probe the influence of CSPs.

In summary, within the scope of this research, intelligibility measured the degree to which

L2 listeners can accurately recognise lexical items (in connected speech vs. citation form) produced by L1 speakers. Comprehensibility evaluated the level of ease or difficulty perceived by L2 listeners in processing connected speech produced by L1 speakers. Detailed descriptions of the corresponding linguistic measures are provided in the subsequent sections on connected speech dictation, citation form dictation, and comprehensibility rating.

3.3 Linguistic Measures

To achieve the objectives of identifying problematic CSP categories in Phase 1 and evaluating the effectiveness of pedagogical interventions in Phase 2, this research implemented an assessment framework comprising six distinct linguistic measures – three principal measures and three supplementary ones. These measures were programmed onto Gorilla Experiment Builder[17] (Anwyl-Irvine et al. 2020, henceforth Gorilla) and undertaken by participants online using their own computer devices.

3.3.1 Principal Measures

Connected Speech Dictation

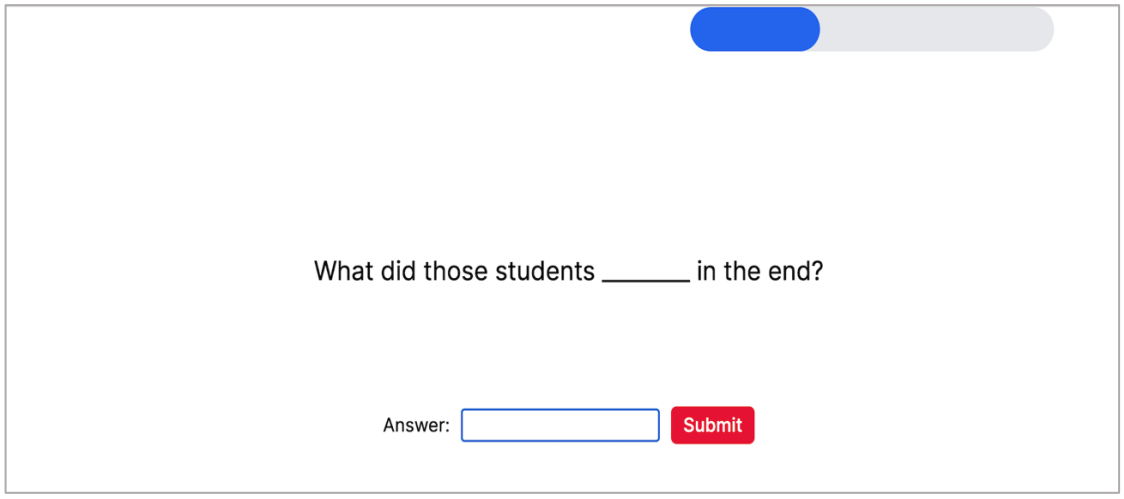
In this task, participants encountered a series of independent sentences, each containing a target CSP. In each trial, they saw the sentence written with a gap where CSP-affected words occurred; before the task began, participants were informed that each gap would represent at least two words missing rather than a single word. After a brief visual presentation (500 milliseconds) of the sentence, an audio recording of the sentence was played which used natural connected speech. Presenting the visual stimuli briefly before

[17]Gorilla Experiment Builder (https://gorilla.sc/) is an online experiment platform developed primarily for research purposes.

the audio prevented participants from being 'caught off guard' in one-off trials. Participants then filled in the blanks with words in standard orthography. To ensure that participants focused on 'perception' first, they could only begin typing their responses after the audio was played in its entirety. Figure 3.2 shows a sample response screen where *find out* are the intended words (containing *CC-V Linking*). The stimulus items were presented in a random order to each participant, and they had 15 seconds to respond to each item. Responses were scored on a binary scale: 1 for accurate answers, 0 for inaccurate ones. Allowances were made for spelling and typographical errors that did not compromise the phonemes involved in the target CSPs (see Appendix A for complete scoring criteria).

Figure 3.2

Linguistic Measure: Connected Speech Dictation



Note. The audio recording of each connected speech item is played with its sentential context presented visually. The response box appears after the audio has finished.

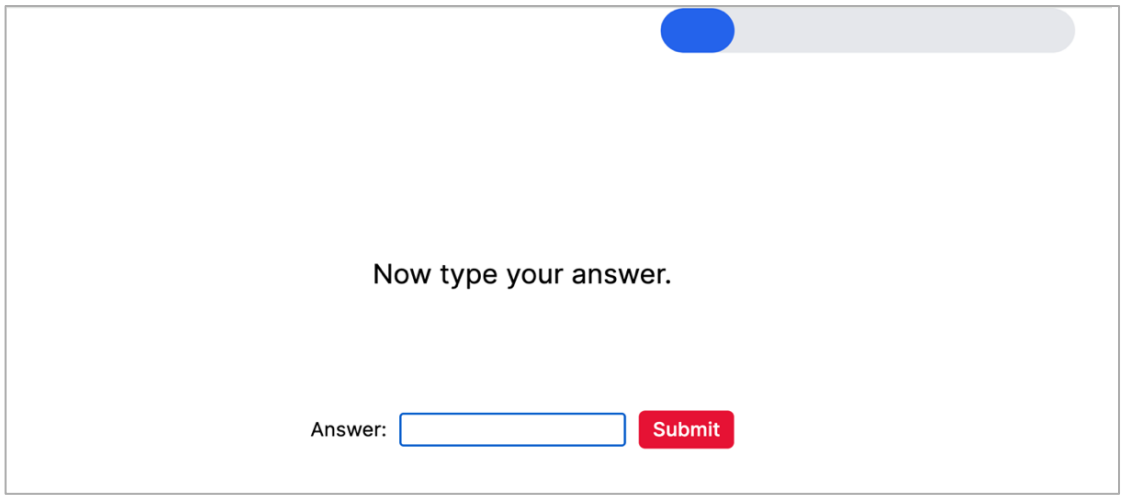
Citation Form Dictation

In this task, participants typed out words they heard in the audio recording. These items were identical to those blanked out in the connected speech dictation but were pronounced

in their citation forms with no sentential context provided either visually or aurally. Participants could only begin typing their responses after the audio finished. Using the same example as above, participants would hear the two words *find* and *out*, separately and each in their citation form. Figure 3.3 shows a sample response screen of this task. Within each item, a fixed silence of 500 milliseconds was placed at the beginning of the audio, between each word, and at the end to ensure these words were perceived individually in isolation, as opposed to their realisation in connected speech. This helped further promote the perception of these items as discrete lexical entries, even though they were already deliberately produced in their citation forms during the stimulus recording process (see Section 4.1.3: *Audio Recording*). The stimulus items were presented in a random order to each participant, and they had 15 seconds to respond to each item. Each item was scored on a binary scale: 1 for accurate answers, 0 for inaccurate ones. The same scoring criteria (Appendix A) applied as those of the connected speech dictation. Discrepancies between the citation form dictation and connected speech dictation scores were used to determine the level of perceptual difficulty attributable to CSPs (participants' familiarity with the target lexical items is addressed in the *L1 Lexical-Semantic Knowledge Test* (LSK) below).

Figure 3.3

Linguistic Measure: Citation Form Dictation



Note. The audio recording of each citation form item is played with no sentential context provided. The response box appears after the audio has finished.

Comprehensibility Rating

In this task, participants rated the comprehensibility of each connected speech item by selecting a response on a 7-point Likert scale[18], as shown in Figure 3.4. Each response represented a different level of ease or difficulty in understanding the item, ranging from 'impossible to understand' at one end, to 'effortless to understand' at the other, with a neutral midpoint denoting 'neither easy nor difficult'. To minimise any confusion regarding scale orientation, visual aids accompanied the textual descriptors: a frowning face represented the 'most difficult' end of the scale, while a smiling face represented the 'easiest' end. This task was administered in Phase 2 alongside the connected speech dictation to assess whether connected speech was perceived to be easier or more difficult to understand after targeted interventions. Participants first heard the connected speech item, then submitted a response for the connected speech dictation (as in Figure 3.1), and

[18]Isaacs and Thomson (2013) found that some raters tend to have difficulty reliably distinguishing all categories on a 9-point scale, particularly in the middle range. Therefore, a 7-point scale with textual descriptors for each point was used in this study to support L2 listeners' judgments.

immediately afterwards they rated the comprehensibility of the same item (as in Figure 3.4). Participants had 10 seconds to respond to each item, and their ratings were coded numerically as an ordinal variable for analysis.

Figure 3.4

Linguistic Measure: Comprehensibility Rating

Note. In Phase 2, the comprehensibility rating screen appears immediately after the submission of a response for the connected speech dictation.

3.3.2 Supplementary Measures

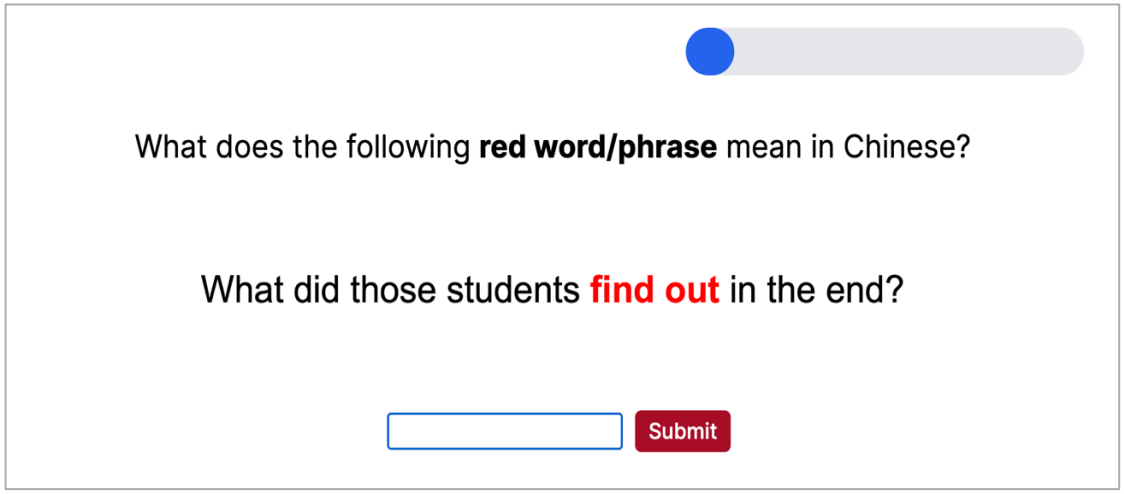
L1 Lexical-Semantic Knowledge Test (LSK)

In the LSK, participants were required to provide the meaning of the target L2 (English) words in their L1 (Mandarin Chinese) upon seeing the visual presentation of complete sentences (i.e. the stimulus items used in the connected speech dictation and comprehensibility rating) with the target words highlighted in red (Figure 3.5). This test aimed to verify that participants' receptive vocabulary knowledge was sufficient, ensuring that any difficulty they experienced in recognising the target words in the connected speech dictation could be reasonably attributed to CSPs rather than inadequate lexical-

semantic knowledge (i.e. not knowing these words even when presented visually in context).

Figure 3.5

Linguistic Measure: L1 Lexical-Semantic Knowledge Test (LSK)



Note. Phrasal items, such as *find out* in this example, are tested as a whole. For non-phrasal items, individual words are tested separately.

If the target words within a stimulus item constitute a phrasal expression, where the meaning of the phrase differs from the combined meanings of its individual components, participants would respond to the phrase as a single semantic unit rather than to each of the constituent words separately. For example, as illustrated in Figure 3.5, the phrase *find out* within the given sentential context elicited one single response from participants rather than two separate responses for *find* and *out*. The stimulus items were presented in a random order to each participant, and they had 15 seconds to respond to each item. Each item was scored on a binary scale: 1 for accurate answers, 0 for inaccurate ones. Allowances were made for typographical slips, such as the incorrect selection of homophonous Chinese characters, provided the intended meaning remained clear. In cases where an item could correspond to more than one possible translation, only

contextually appropriate meanings were accepted.

It is important to note that during the devising of stimuli, advanced vocabulary was deliberately excluded, with over 70% of the words classified at the beginner level (Table 3.2). Given this, it was anticipated the learner group – whose English proficiency ranged from upper-intermediate (CEFR B2) to advanced (CEFR C1/C2) levels (refer to Phase 1 and Phase 2 participant profiles in Sections 4.1.1 and 5.1.1, respectively) – would be able to recognise the majority of the target words. Therefore, a high, or potentially ceiling-level accuracy rate was expected for this measure.

Table 3.2

The Level of Target Lexical Items in Experimental Stimuli

CEFR	**Beginner**		**Intermediate**		**Advanced**			**Total**
Level	A1	A2	B1	B2	C1	C2	n/a	
Words	104	57	30	18	0	0	8	217
Percentage	47.9%	26.3%	13.8%	8.3%	0%	0%	3.7%	100%

Note. The level of each lexical item is based on English Vocabulary Profile Online established by Cambridge University Press and Assessment (2025). Items marked as 'n/a' indicate words or phrases for which no CEFR level was available in the database.

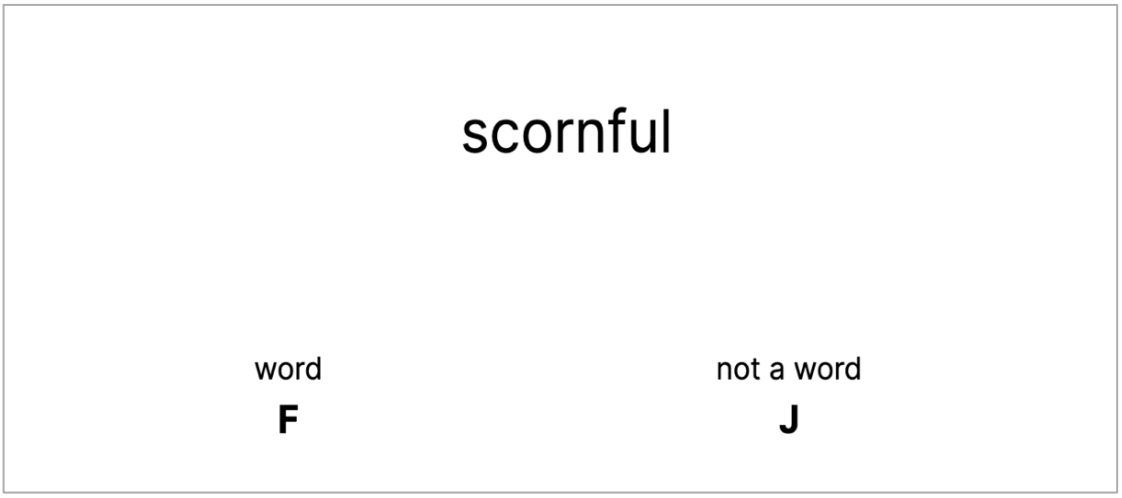
Lexical Test for Advanced Learners of English (LexTALE)

Originally developed by Lemhöfer & Broersma (2012), the LexTALE is a lexical decision task which correlates significantly with advanced learners' overall proficiency in English. In the LexTALE, participants saw independent letter strings one at a time and decided whether each letter string was an existing English word or a nonword. Participants

responded to each item within a two-second timeframe by pressing on the keyboard either ‘F’ for a word or ‘J’ for a non-word. A buffer time period of 500 milliseconds was inserted as a transition between each item. Figure 3.6 shows a sample response screen. The whole test comprised 60 items (40 words and 20 nonwords), each scored on a binary scale: 1 for accurate answers, 0 for inaccurate ones. The overall accuracy was calculated based on the average of word accuracy and nonword accuracy – for example, if 50% of the existing words were accurately identified and 70% of the nonwords were accurately rejected as words, the overall accuracy would be 60%.

Figure 3.6

Linguistic Measure: Lexical Test for Advanced Learners of English (LexTALE)



The LexTALE served three purposes in this research. First, it provided a proxy measure of L2 proficiency for participants who did not possess scores from any standardised proficiency tests. Second, it offered a consistent proficiency scale for eligible participants who took various standardised proficiency tests, some of which may not be directly comparable and only indicate approximate proficiency ranges. Third, it enabled more nuanced analyses by allowing proficiency to be treated as a continuous variable, even within broadly defined proficiency categories.

L2 General Listening Comprehension Test (GLC)

Adapted from the Listening Module of the International English Language Testing System (IELTS), the GLC assessed participants' overall L2 listening comprehension. This aligned with the recommendation by Cross and Vandergrift (2015) that 'researcher-designed tests that target a particular type of listening input and behaviour should be complemented by standardized tests (for example IELTS)' (p. 88). To accommodate participants' attention span and cognitive load, given the number of tasks involved in the experiment (as described above), a shortened version of the IELTS Listening[19] test was employed. Specifically, the GLC included only two passages instead of a full set of four. The full IELTS Listening Module typically consists of two monologues and two conversations, with each type covering one social scenario and one academic scenario. For the purposes of this research, conversational passages were deemed more appropriate, as monologues (e.g. lectures, presentations, and announcements) tend to be delivered in a more careful speech style, less likely to be influenced by CSPs. The GLC thus comprised two conversational passages, one in a social and one in an academic scenario. Participants responded to 10 comprehension questions per passage, totalling 20 questions. Scores were calculated as a percentage of correct answers out of 20. To ensure test validity, upon completion of the task, participants were explicitly asked whether they had previously encountered any of the materials; if any prior exposure was reported, the relevant scores were excluded from analysis. This measure allowed for the exploration of potential correlations between L2 general listening comprehension and the intelligibility and comprehensibility of connected speech as perceived by L2 learners.

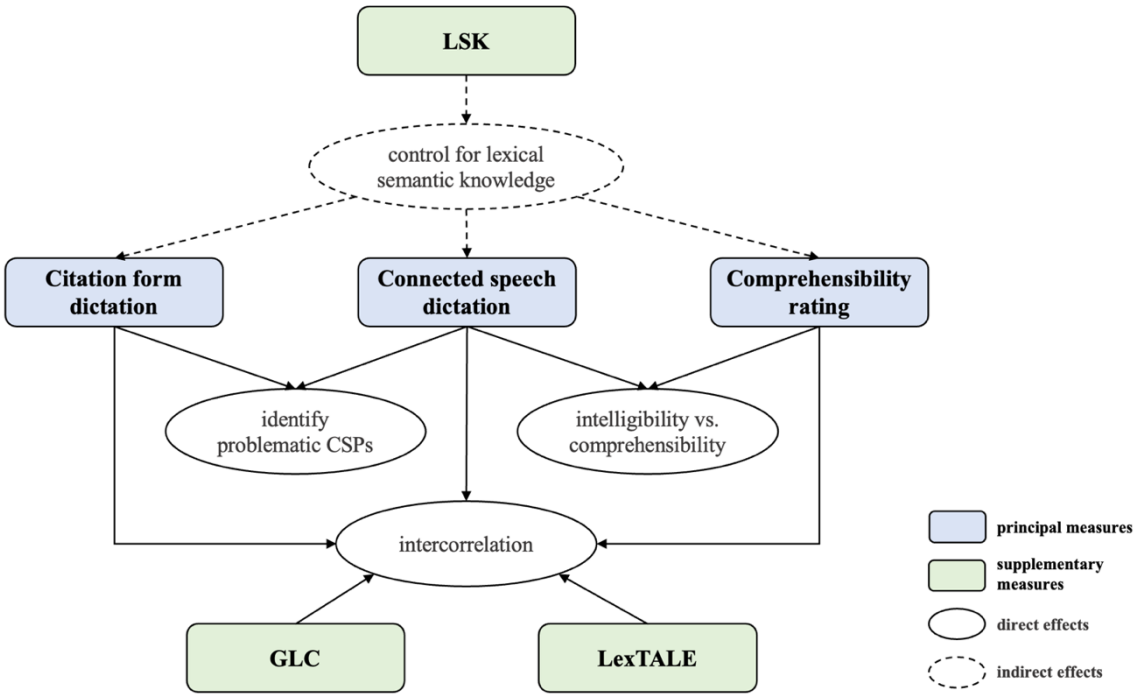
[19]The listening passages used for assessment were adapted from the *Cambridge IELTS Academic Authentic Practice Tests* series published by Cambridge University Press and Assessment.

Summary

The Phase 1 study was designed to examine the impact of English CSPs – and the relative impact of different categories of CSP – on L2 learners' speech perception. The methodology focused on analysing the performance gap between the connected speech and citation form dictation tasks, both of which utilised identical lexical items to isolate the effects of CSPs. The subsequent Phase 2 study was designed to evaluate the effects of pedagogical interventions on two main outcomes: (i) changes in the ability to recognise words affected by CSPs in connected speech (as measured by transcription accuracy-based intelligibility scores) from the pre-test to post-test; and (ii) changes in the perceived ease or difficulty in understanding connected speech (as measured by comprehensibility ratings). Additionally, changes in general listening comprehension (as measured by the GLC) may potentially provide insights into broader pedagogical effects. In both phases, the LSK and LexTALE were used to control for participants' L1 lexical-semantic knowledge and overall L2 proficiency, respectively. Intercorrelations among the principal and supplementary measures may provide a clearer perspective on their dynamic interplay and help to contextualise the interpretation of findings. Figure 3.7 illustrates the structural relationships among the six linguistic measures employed in this research.

Figure 3.7

Structural Relationships among the Employed Linguistic Measures



On the basis of the methodology laid out above, the subsequent two chapters delve into the details of the Phase 1 and Phase 2 studies, respectively. Each chapter provides a comprehensive elaboration of the methods and corresponding experimental results.

3.4 Ethical Considerations

This research project (reference number: CIA-22HT-067) received ethical approval from the Central University Research Ethics Committee (CUREC) at the University of Oxford. Participation was restricted to adults aged 18 years or over. In both phases of the research, informed consent was obtained from all participants prior to the commencement of data collection, ensuring that they understood the study's purpose, procedures, and their right to withdraw from the study at any time without consequences.

In addition to their responses in the experiments, age and gender were collected as

demographic information. Participants were identified with pseudonyms; the key linking pseudonyms to their email addresses was stored separately from the research data and destroyed at the end of the study.

All data management and protection procedures adhered to the *Best Practice Guidance 09* issued by the CUREC and the *Ethical Guidelines for Educational Research* by the British Educational Research Association (BERA, 2018).

Chapter 4: Phase 1 Study – Diagnosis

The Phase 1 study was designed to examine the impact of CSPs and to identify challenging CSP categories for adult Mandarin-speaking learners of L2 English. This phase involved selecting CSP categories for investigation, devising a stimulus inventory, and conducting a diagnostic experiment with the target participants. This chapter first details the methods employed in Phase 1 and subsequently presents the results of the experiment.

4.1 Methods

4.1.1 Participants

The Phase 1 study targeted adult learners of English who speak Mandarin as their L1 and have achieved an advanced proficiency level in English as their L2, corresponding to C1 level or above on the CEFR[20] scale (see Appendix B for the comparison of scores for each level in common standardised proficiency tests). To ensure participants met this criterion, they were asked to provide their most recent score on a standardised English proficiency test and took the LexTALE as an additional indication of their proficiency level (cf. Supplementary Measures: *Lexical Test for Advanced Learners of English* (LexTALE) for task details).

A total of 50 adult Mandarin-speaking learners of English (30 female, 20 male) were

[20]The Common European Framework of Reference for Languages (CEFR, https://www.coe.int/en/web/common-european-framework-reference-languages) is an international guideline for classifying language proficiency into six levels: A1 – A2 for basic users, B1 – B2 for independent users, C1 – C2 for proficient users.

recruited[21]. Initially, three participants reported either an unknown or borderline proficiency level[22]. However, they were later included based on their scores in the three supplementary measures: LexTALE (overall proficiency), GLC (L2 general listening comprehension), and LSK (L1 lexical-semantic knowledge test for the target words). To pass the inclusion thresholds, their scores needed to exceed the cut-off point set at two standard deviations below the mean values of the other 47 certified advanced participants. This cut-off point approximated the lower bound for including 95% of the data in a normal distribution (though not assuming a perfect normal distribution in the data obtained). This approach ensured that their scores fell within the expected range for advanced L2 proficiency and demonstrated comparable linguistic abilities to the other certified participants. As displayed in Table 4.1 below, these three participants' scores in the LexTALE, GLC, and LSK all surpassed the thresholds, thereby confirming their eligibility for participation in the experiment.

Table 4.1

Inclusion Criteria for Participants with Unknown or Borderline Proficiency

	LexTALE	**GLC**	**LSK**

[21]The sample size of 50 learners was determined with reference to previous work investigating CSP-related challenges. For example, Henrichsen (1984) tested 50 L2 learners and 15 L1 speakers, while Ito (2001) included 18 L2 learners and 9 L1 speakers. Both studies found clear patterns in L2 learners' perceptual difficulty with CSPs (cf. Section 2.3.2 for a review of these studies), supporting the use of the present sample in the Phase 1 diagnostic study.

[22]One participant (M2) passed the *Test for English Majors – Band 8* and *China Accreditation Test for Translators and Interpreters – Level 2*, both of which were not directly transferrable to the CEFR scale; another participant (M30) reported an overall band score of 6.5 in the IELTS, which was considered the borderline level between B2 and C1; the other participant (M73) reported that they had never taken any standardised English proficiency test before.

Mean	68.5%	75.6%	93.9%
Standard Deviation (SD)	11.3%	11.6%	3.2%
Threshold (Mean – 2*SD)	45.9%	52.4%	87.5%
Participant M2	51.3%	75.0%	96.8%
Participant M30	95.0%	80.0%	92.5%
Participant M73	82.5%	85.0%	96.2%

Additionally, 20 adult native English-speaking participants (10 female, 10 male) were recruited as a comparison group. They were expected to demonstrate a robust ability to decode connected speech, with minimal impact from the presence of CSPs.

4.1.2 CSP Categories Investigated

Connected speech can vary considerably depending on factors such as phonetic or phonological features in different language varieties and idiosyncratic patterns of individual speakers. Furthermore, variations at different levels (e.g. phonemic, lexical, and phrasal) may co-occur with each other, given the dynamic nature of connected speech. This means in any utterance there exists an abundance of features, subtle or overt, that may be considered the result of CSPs. Therefore, to judiciously select CSP categories for investigation, five inclusion/exclusion criteria were established. Specifically, the following criteria had to be met for a CSP category to qualify for inclusion in this research:

Criterion 1: Inter-Word Phonological Environment

This study specifically investigated CSPs which occur in the inter-word phonological environment (i.e. across word boundaries). That is, the CSP mechanisms simultaneously involve the final segment(s) of one word and the initial segment(s) of the following word,

typically spanning two lexical items but occasionally extending to three (as in the *Multiple* category). As such, processes that occur within single words, such as word-internal insertion (e.g. *prince* realised as /prɪnts/), syncope (e.g. *family* realised as /ˈfamli/), and aphesis (e.g. *about* realised as *'bout*) were excluded.

Criterion 2: Non-Lexicalisation

Contracted forms with standardised orthographic representations (e.g. *haven't* for *have not*) were excluded, since their usage and pronunciation are highly lexicalised. In pedagogical contexts, these forms are more typically treated as distinct lexical items alongside their non-contracted equivalents, rather than being attributed to phonological alterations in spontaneous connected speech.

Criterion 3: Universality

CSPs that are primarily characteristic of specific English varieties were excluded. Examples include alveolar flapping generally found in General American English (e.g. *put on* realised as /pʊɾɑn/[23]) and intrusive *r* in certain British English varieties (e.g. *idea of* realised as /ʌɪˈdɪəɹəv/). This criterion was applied to ensure that this study focused on CSP categories that are more universally attested across major varieties of English.

Criterion 4: High Frequency

Low-frequency CSPs were also excluded from this study. For instance, dissimilation (e.g. *fifths* realised as /fɪfts/) occurs infrequently in English and as noted by Celce-Murcia et al. (2010), can be 'ignored for pedagogical purposes' (p. 171). This focus allowed the

[23] An inter-word alveolar flapping typically has a vowel-consonant-vowel environment and is thus categorised under *VC-V Linking* in this research.

current study to prioritise CSP categories with greater relevance and applicability to language teaching and learning contexts, rather than those that may be rarely encountered in natural speech and therefore less likely to impede real-life speech perception.

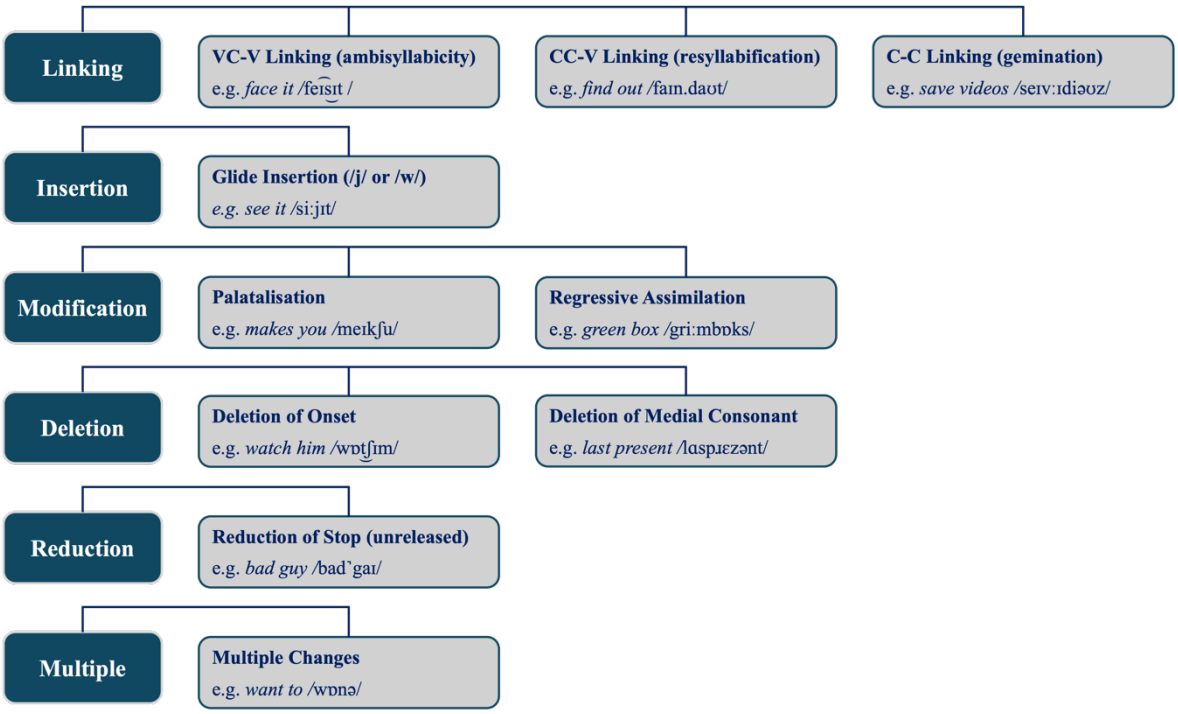
Criterion 5: High Saliency

Lastly, CSPs with particularly low perceptual saliency – i.e. those that are acoustically difficult to detect by ear – were excluded from the study. For example, delayed release (e.g. *about the* realised as /əˈbaʊ.t̚ðə/) is almost indistinguishable from non-release (Hieke, 1987a) and can therefore be considered to have minimal perceptual impact on listeners. This criterion helped to ensure that the selected CSP categories were sufficiently audible, thereby enhancing their relevance for pedagogical attention.

Based on the criteria outlined above, ten categories of CSPs as defined in the classificatory scheme proposed by Alameen and Levis (2015), grouped into six overarching categories, were targeted for investigation (Figure 4.1). These categories were *VC-V Linking*, *CC-V Linking*, *C-C Linking*, *Glide Insertion*, *Palatalisation*, *Regressive Assimilation*, *Deletion of Onset, Deletion of Medial Consonant*, *Reduction of Stop*, and *Multiple*. For detailed phonological mechanisms underlying each category, refer to the review of *Phonological Mechanisms of Common CSPs in English* in Section 2.1.2.

Figure 4.1

CSP Categories under Investigation



4.1.3 Devising of Stimulus Inventory

A total of 116 experimental stimulus items were devised, each containing CSP-affected words embedded within a sentential context. Between 10 and 15 items were devised for each of the ten CSP categories, allowing for a balanced representation of viable phonological contexts (to be discussed in detail below in *Considerations of Phonological Contexts in Stimuli*). In addition, 58 distractor items were developed in which no investigated CSPs occurred at the blanked-out position in the connected speech dictation (cf. Section 3.3.1 Principal Measures). These distractors were included to prevent participants from discerning systematic patterns in CSP occurrence. The final ratio of experimental stimuli to distractors was 2:1, with distractors comprising one-third of the total set. This proportion was judged to be sufficient to obscure the distribution of CSP categories across the stimuli, particularly given that all items were pooled and randomised together for presentation. Furthermore, to control participants' cognitive load taxed by

processing each item, the sentential context of each stimulus item was limited to a maximum of 12 syllables (mean = 10.77, SD = .98), aligning approximately with the length at which a pause occurs in natural speech (Field, 2003). A full list of the stimulus items is provided in Appendix C.

Considerations of Phonological Contexts in Stimuli

The CSP-affected lexical items within the experimental stimuli were judiciously designed to ensure a balanced representation of viable phonological contexts – i.e. the constituent segments involved in each CSP category. Because each CSP is triggered by specific phonological environments, distinct strategies were employed in the selection of target segments for each category. These considerations are outlined below:

VC-V Linking. In this category, a word-final consonant that is preceded and followed by a vowel becomes ambisyllabic, functioning simultaneously as the coda of the first word and the onset of the following word. A total of 10 experimental stimuli were devised, each featuring a different consonant within the VC-V sequence. The target consonants included 3 stops, 3 fricatives, 2 affricates, and 2 nasals.

CC-V Linking. In this category, the second consonant within a word-final two-consonant cluster is resyllabified as the onset of the following word. A total of 14 experimental stimuli were devised, each featuring a different consonant cluster within the CC-V sequence. The target resyllabified consonants included 6 stops, 6 fricatives, and 2 affricates.

C-C Linking. In this category, two identical consonants straddling a word boundary

(i.e. one as the coda of a word, the other as the onset of the following word) undergo articulation lengthening, rather than being pronounced twice separately. A total of 12 stimuli were devised, each featuring a different set of geminate consonants. The target consonants included 6 stops, 4 fricatives, and 2 nasals.

Glide Insertion. In this category, a glide is inserted between two vowels across a word boundary (i.e. one word ends with a vowel and the following word begins with a vowel). The glide /j/ is inserted if the preceding word-final vowel is close front (/i/ or /ɪ/), and the glide /w/ is inserted if the preceding word-final vowel is close back (/u/ or /ʊ/). A total of 12 stimuli were devised, 6 for the insertion of /j/ and 6 for the insertion of /w/. Among these items, half of the word-final vowels were monophthongs (i.e. /i/ or /u/), and half were the off-glides of diphthongs (i.e. /aɪ/, /eɪ/, /ɔɪ/, /aʊ/, or /əʊ/).

Another layer of balance in this category was considered in terms of orthographic representation. Some target words contain letters <y> or <w> as part of the word final vowel where the glide insertion occurs (e.g. *say it* and *grew up)*, which might prime participants due to strong sound-spelling correspondences between /j/ and <y> and between /w/ and <w>. To account for this potential priming effect, half of the items – 3 for the insertion of /j/ and 3 for the insertion of /w/ – involved words with such orthographic cues (e.g. *say it*, *grew up*), and the other half without such orthographic cues (e.g. *be able*, *blue ink*).

***Palatalisation*.** In this category, if alveolar fricatives /s/, /z/, stops /t/, /d/, or their combinations /ts/, /dz/ at the word-final position are followed by a word-initial palatal glide /j/, then these alveolar segments are palatalised and become post-alveolar fricatives

/ʃ/, /ʒ/ or affricates /tʃ/, /dʒ/. A total of 15 stimuli were devised, 5 for alveolar stops /t/ or /d/, 4 for alveolar fricatives /s/ or /z/, 4 for the combinations /ts/ or /dz/, and the other 2 involved high-frequency interrogative phrases *would you* and *could you*.

Regressive Assimilation. In this category, a word-final alveolar nasal /n/ assimilates to the place of articulation of the following word-initial stop. When followed by labial stops (/p/ or /b/), /n/ becomes /m/; when followed by velar stops (/k/ or /g/), /n/ becomes /ŋ/. A total of 10 stimuli were devised, 5 for an alveolar /n/ assimilated to a labial /m/, and 5 for an alveolar nasal /n/ assimilated to a velar /ŋ/.

Deletion of Onset. In this category, the onset consonant of pronominal forms or auxiliaries is deleted, particularly when it is *not* stressed or at the beginning of an utterance. A total of 12 stimuli were devised: 7 items involved third-person singular pronominal forms with a glottal fricative onset /h/, including the subject form (*he*), object forms (*him*, *her*), possessive adjectives (*his*, *her*), and possessive pronouns (*his*, *hers*); 2 items involved the third-person plural object with an interdental fricative onset /ð/ (*them*); 3 items involved auxiliaries with a glottal fricative onset /h/ (*have*, *has*, *had*).

Deletion of Medial Consonant. In this category, the consonant in the medial position with an inter-word three-consonant cluster (i.e. the CC-C sequence) is deleted. Two frequently deleted segments are alveolar stops /t/ and /d/ (cf. Section 2.1.2: *Deletion*). A total of 10 stimuli were devised, 5 for the deletion of /t/ and 5 for the deletion of /d/.

Reduction of Stop. In this category, a word-final stop is unreleased when the following word begins with a stop or an affricate. A total of 10 stimuli were devised, with

5 involving stop-stop sequences and 5 involving stop-affricate sequences. The target unreleased segments included all six stops in English (/p/, /t/, /k/, /b/, /d/, /g/).

Multiple. In this category, high-frequency phrasal expressions undergo multiple phonological alterations. A total of 11 such stimuli were devised. Among these, 3 items featured phrases with casual but conventional orthographic representations (e.g. *wanna* representing *want to*); although these items may appear to contravene the *Non-Lexicalisation* criterion (cf. Section 4.1.2), these forms are generally considered informal and primarily used in spoken language rather than fully lexicalised contractions like *doesn't* or *haven't*. The remaining 8 items featured phrases with informal and non-conventional orthographic representations (e.g. **useta* representing *used to* and **Whaddaya* representing *What are you* or *What do you*), reflecting phonological and colloquial variants common in spontaneous connected speech.

Piloting of Stimuli

Following the initial devising of stimulus items, two pilot experiments were conducted with native English speakers to ensure the appropriateness of each item.

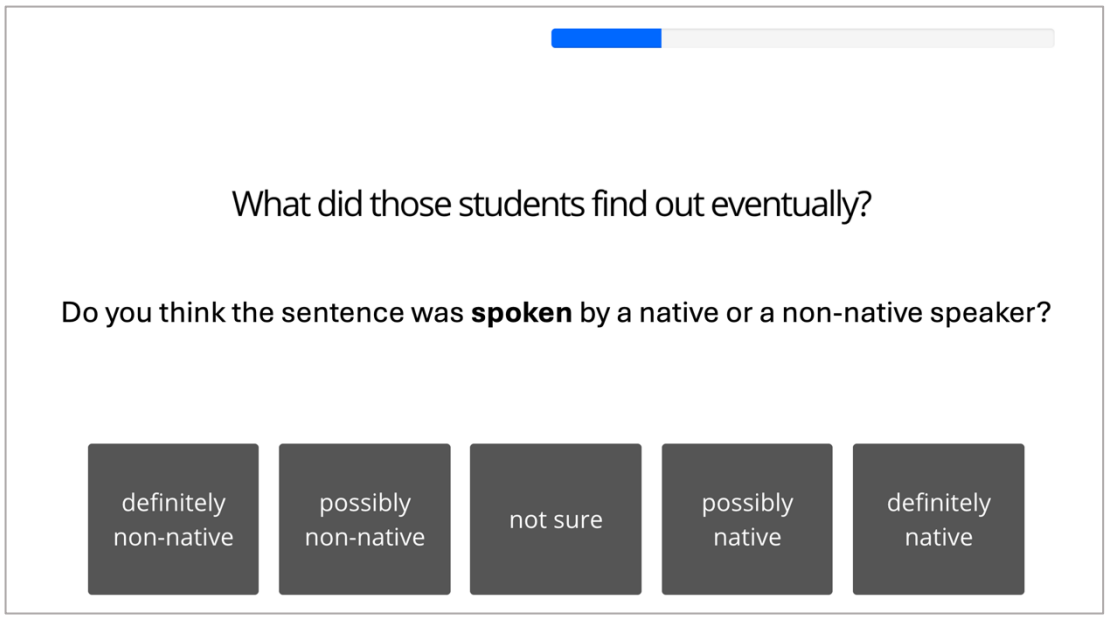
Grammaticality Pilot Experiment. The first pilot experiment focused on the grammaticality[24] of individual stimuli. Four native English speakers participated as informants, with two assigned to each split-half set of stimuli. They completed a rating

[24]The term 'grammaticality' used in this pilot concerns whether a sentence is both accurate (i.e. free of grammatical errors) and natural in *spoken* English (i.e. native-like usage), given that non-native-like utterances may result from ungrammatical and/or unnatural expressions. This bears relevance to the fact that greater naturalness in the stimuli facilitates the use of CSPs, as it enables more natural speech production, particularly during audio recording.

task in which they read the full sentences and evaluated each item on a 5-point Likert scale, indicating how likely each item was spoken by a native or a non-native speaker. A sample response screen is shown in Figure 4.2. If an informant rated a sentence as 'definitely non-native', 'possibly non-native' or 'not sure', they were asked to provide a revised version that sounded more native-like to them. Consensus among informants was achieved for 72.8% of the stimuli. For non-consensual items – i.e. those receiving conflicting judgements – revisions were made in consultation with two faculty experts in applied linguistics who are native speakers of English.

Figure 4.2

Grammaticality Pilot Experiment



Predictability Pilot Experiment. The second pilot experiment assessed the predictability of the target words containing CSPs within the sentential context of each experimental stimulus. This experiment was designed to ensure that even native informants were unable to reliably predict the target words based solely on the syntactic

and pragmatic context of each sentence. This control served to prevent participants from being able to consistently predict the target words without having to attend to the auditory input during the connected speech dictation (i.e. by leveraging top-down processing; see the review of *Top-Down and Bottom-Up Processing* in Section 2.2.3). Otherwise, the validity of the dictation measure – designed to evaluate participants' ability to identify CSP-affected words from the auditory input – could be compromised.

Each round of testing involved five native English-speaking informants, with a total of sixteen[25] informants participating across three rounds of testing. Informants were shown the stimulus sentences with the target words blanked out and were requested to complete each item by filling in those blanks based on their intuition from merely reading it, without access to the audio. The visual presentation and response format of this pilot experiment mirrored those of the connected speech dictation measure (cf. Figure 3.2), except that no auditory input was provided.

Stimulus items were revised and retested if they were either fully predicted (i.e. the participant's response matched both intended target words) or partially predicted (i.e. the participant's response matched only one, but not both, intended target words) by three or more out of five informants. The testing was conducted iteratively over three rounds until no experimental stimuli could be consistently predicted. In Round 1, all 116 experimental stimuli were tested, of which 26 items met the predictability threshold and were revised and retested in Round 2. Following that, 8 of the items required further revision and were

[25]Ten informants participated in the first round, with each group of five assigned to a split-half stimulus set. In both the second and third rounds, five informants participated in each. Among all the informants involved across the three rounds of predictability testing, four participated in two different rounds, each time responding to a different subset of items.

retested again in Round 3. The final results of the predictability testing are summarised in Table 4.2. Eventually by the end of the third round, 90.5% (n = 105) of the stimuli elicited 0 correct prediction, 9.5% (n = 11) elicited only 1 correct prediction, and no item elicited more than 1 correct prediction from the five native informants.

Table 4.2

Results of Predictability Pilot Experiment

	Prediction Count for Experimental Stimuli (N = 116)			
	Prediction = 0	Prediction = 1	Prediction > 1	Sum
Experimental Stimuli	105	11	0	116
Percentage	90.5%	9.5%	0%	100%

Audio Recording

Upon the completion of the two pilot experiments, the stimuli were audio recorded in a soundproof laboratory by six native speakers of English (three females and three males) who speak general mid-southern UK English[26]. For each stimulus item, both connected speech and citation form versions were recorded by the same speaker in order to control for speaker-related variability. This ensured that any discrepancies in participants' recognition of target words across the two conditions could be attributed to the presence or absence of CSPs, rather than to differences in speaker identity or voice quality. Detailed audio recording procedures and instructions provided to the audio contributors are

[26]The English varieties spoken by the native audio contributors were identified based on two criteria: (1) the speakers' self-reported hometowns where they had spent the majority of their lives; and (2) expert judgements by two applied linguists – both native speakers of UK English – who evaluated sample audio recordings by each contributor.

documented in Appendix D.

Following the preliminary recording, a phonetic analysis was conducted using Praat[27] (Boersma & Weenink, 2024) to examine the spectrographic properties of the audio stimuli. In particular, formant structures – specifically the first three formants (F1, F2 and F3) – were inspected for the segments affected by target CSPs, to ensure a clear presence of the intended phonological mechanisms. Any items that appeared either spectrographically or aurally ambiguous underwent a second recording before finalisation.

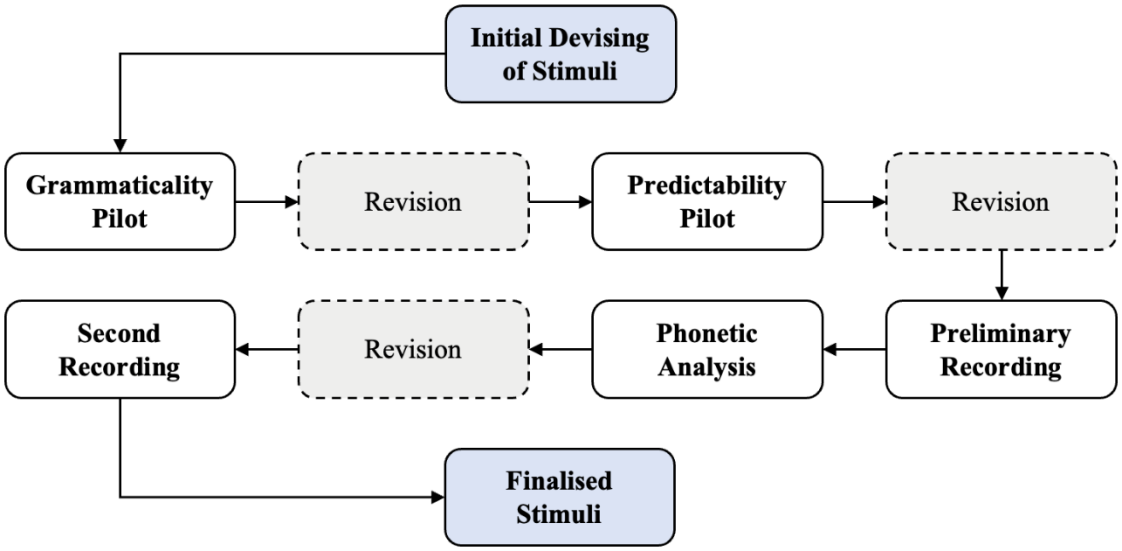
Summary of Stimulus Inventory

A total of 116 experimental stimuli and 58 distractors were developed. The experimental stimuli were strategically designed to ensure balanced phonological contexts within each CSP category under investigation. Two pilot experiments were conducted to evaluate these stimulus items, and the results confirmed that (1) all stimuli were judged to be grammatical and natural in spoken English; and (2) the target words featuring CSPs within the stimuli were completely or largely unpredictable based on their syntactic and pragmatic contexts alone. Following these validations, the stimuli were audio recorded and subjected to a phonetic analysis to verify the presence of the intended CSPs. The overall process of stimulus development is schematically summarised in Figure 4.3.

Figure 4.3

[27] Praat (http://www.praat.org/) is a widely used software tool for phonetic analysis, developed by the Phonetic Sciences Department of the University of Amsterdam.

Process of Stimulus Development



4.1.4 Implementation Procedures

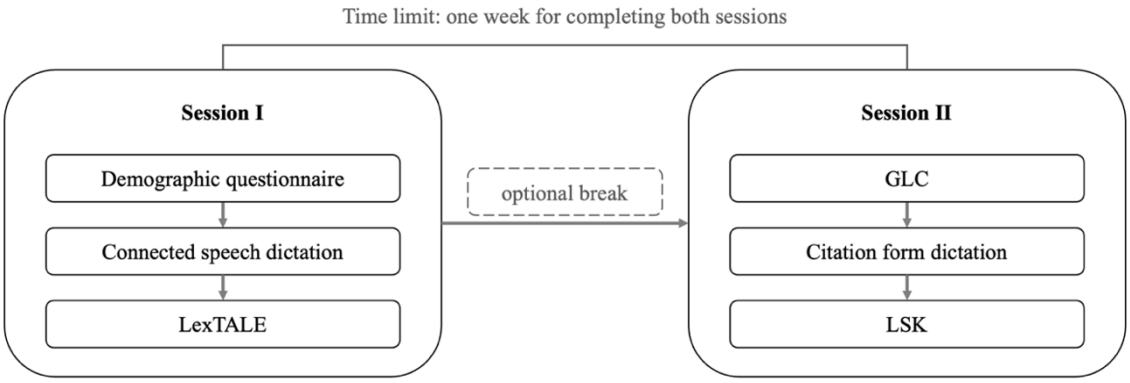
The procedures for the Phase 1 study comprised implementing five linguistic measures: connected speech dictation, LexTALE, GLC, citation form dictation, and LSK (cf. Section 3.3 for the details of these measures). These measures were administered in this sequential order to assess the influence of CSPs. The connected speech dictation preceded the citation form dictation since the latter was generally easier for L2 listeners to process; if the citation form dictation had preceded the connected speech one, then it might have primed participants to recognise the target words affected by the CSPs. The LSK followed the two dictation tasks, again to prevent a priming effect whereby participants could have gained an advantage of knowing what words would be used in the stimuli. The LexTALE and GLC were undertaken as 'buffers' between the other three measures, distracting participants from working repeatedly with the same stimuli.

All the measures along with a demographic questionnaire were programmed into one experiment on Gorilla and completed by participants on their own computer device over

two 45-minute sessions. Participants were offered an optional break after they had completed Session 1 and were required to complete both sessions within one week. Figure 4.4 illustrates the experiment procedures for L2 participants. Native English-speaking participants (serving as a comparison group) completed only the connected speech dictation and citation form dictation to establish norms of word recognition minimally affected by CSPs.

Figure 4.4

Phase 1 Experiment Procedures for L2 Participants



4.2 Results

In the Phase 1 study, L2 participants' performance across five linguistic measures – connected speech dictation, citation form dictation, LexTALE, GLC, and LSK (cf. Section 3.3) – was analysed statistically to determine the extent to which CSPs impact L2 learners' speech perception. In addition, L1 (native) English speakers' performance on the two dictation tasks served to establish a baseline for comparison. The subsequent sections report the results derived from the analyses.

4.2.1 Connected Speech Dictation vs. Citation Form Dictation

The first step of analysis involved calculating each participant's word recognition accuracy across the two dictation measures: connected speech and citation form. Table 4.3 presents the number of data points collected for these two principal measures. Each data point corresponds to one participant's response to one experimental stimulus (of which there were 116 in total). It can be seen that a small number of data points were missing from the L2 group (had there not been any missing data, the number would have been 5,800 per measure), likely due to technical response recording issues during data collection on Gorilla.

Table 4.3

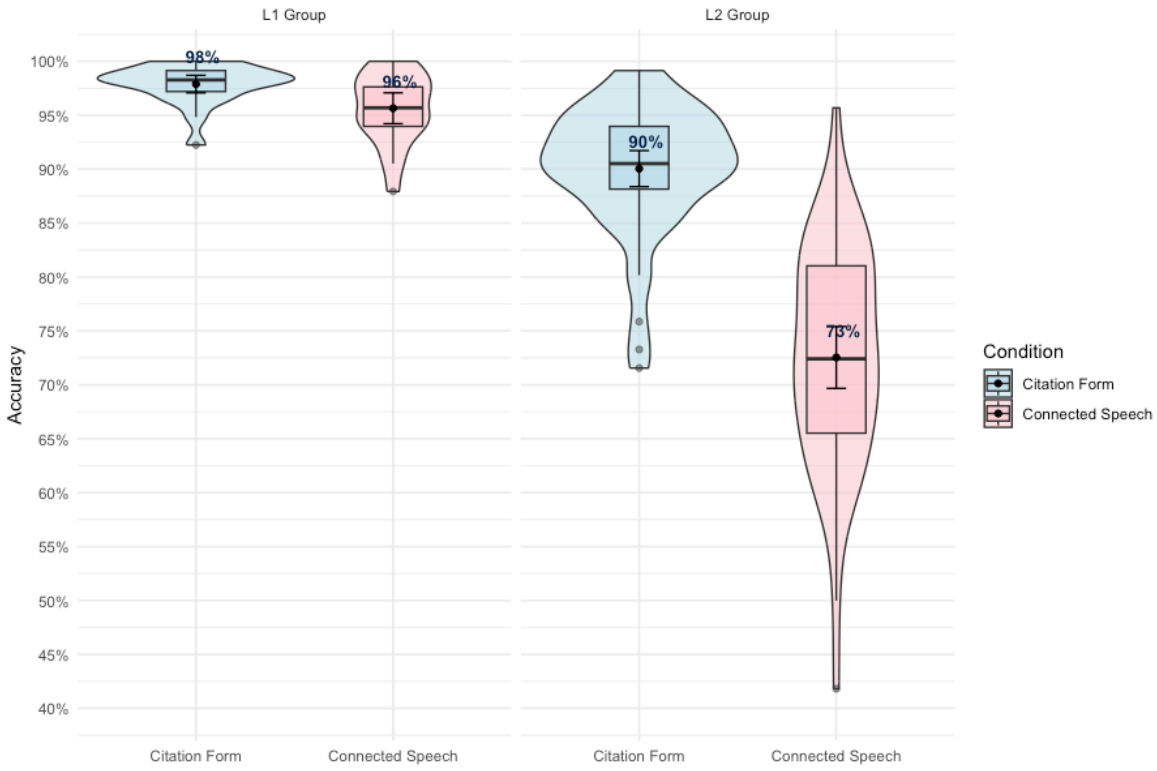
Data Points for Connected Speech Dictation and Citation Form Dictation

	English (L1) Group n = 20	**Mandarin (L2) Group n = 50**	**All Participants N = 70**
Connected speech dictation	2,320	5,770	8,090
Citation form dictation	2,320	5,799	8,119
Total	4,640	11,569	16,209

Figure 4.5 illustrates the distributions of mean accuracy for the two dictation measures using violin plots. The L1 group is shown on the left and the L2 group on the right. As expected, and aligning with the literature (e.g. Henrichsen, 1984; Ito, 2001), the L2 group exhibited substantially lower accuracy in recognising the target words produced with CSPs in connected speech (mean accuracy = 73%) compared to the same words in citation form (mean accuracy = 90%), reflecting a disparity of 17 percentage points. In contrast, the L1 group demonstrated nearly equivalent performance on connected speech (mean accuracy = 96%) and citation form (mean accuracy = 98%), yielding only a marginal disparity of 2 percentage points.

Figure 4.5

Connected Speech versus Citation Form Accuracy by Group



Note. Within each violin plot, the boxplot shows the interquartile range with the bold horizontal line representing the median accuracy. The error bars indicate 95% confidence intervals around the means.

To confirm the observed patterns, a mixed effects logistic regression model[28] was fitted to examine participants' performance as a function of Condition (citation form vs. connected speech), Group (L1 vs. L2), and their interaction. Both predictors were centre coded. The model included random by-participant intercepts and random by-participant slopes for Condition. Results revealed significant main effects of Condition (b = -1.06,

[28]The model was fitted using the *lme4* package in R (Bates et al., 2015).

SE = 0.10, z = -10.41, p < .001) and Group (b = -1.94, SE = 0.18, z = -10.87, p < .001), indicating that overall accuracy was significantly lower in connected speech than citation form, and lower for the L2 group than the L1 group. Crucially, there was a significant Condition x Group interaction (b = -0.51, SE = 0.20, z = -2.60, p = .009), indicating that the L2 group experienced a significantly greater decline in accuracy in response to connected speech compared to the L1 group.

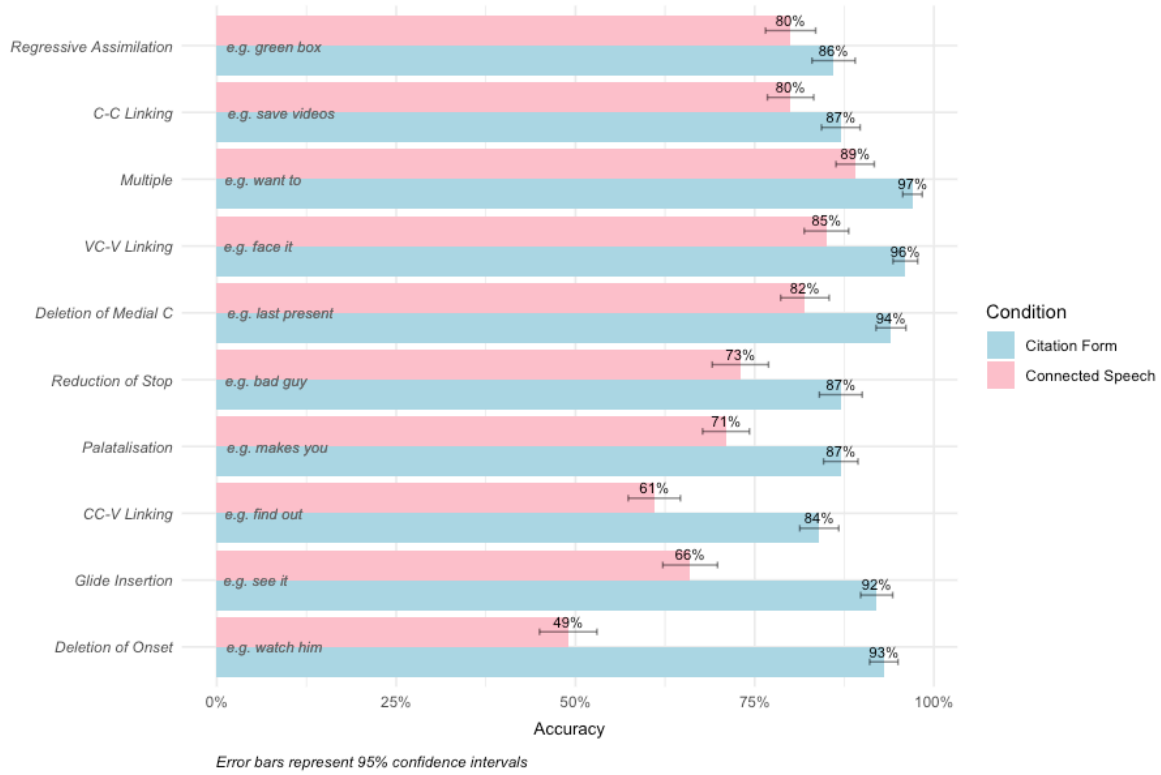
Based on these results, the next step was to further examine the extent to which individual CSP categories contributed to the pronounced disparity in processing words in connected speech versus citation forms among this advanced L2 learner group – i.e. whether different CSP categories exerted similar or varying impacts on their speech perception.

4.2.2 Level of Difficulty in Processing Different CSP Categories

To further explore the L2 learners' difficulties, their scores for the two dictation measures were disaggregated by CSP category. Figure 4.6 displays the mean accuracy for each CSP category in both citation form and connected speech, presented in order of increasing disparity in accuracy between the two conditions: categories with the smallest disparities (i.e. least challenging) are shown at the top and those with the largest disparities (i.e. most challenging) at the bottom.

Figure 4.6

Connected Speech versus Citation Form Accuracy by CSP Category



In this initial analysis, categories with connected speech accuracy below 75% were considered particularly challenging. From the data, five out of ten CSP categories (the bottom five in Figure 4.6) met this criterion which also exhibited the largest disparities: *Deletion of Onset* (49% accuracy in connected speech; a disparity of 44 percentage points), *Glide Insertion* (66% accuracy in connected speech; a disparity of 26 percentage points), *CC-V Linking* (61% accuracy in connected speech; a disparity of 23 percentage points), *Palatalisation* (71% accuracy in connected speech; a disparity of 16 percentage points), *Reduction of Stop* (73% accuracy in connected speech; a disparity of 14 percentage points).

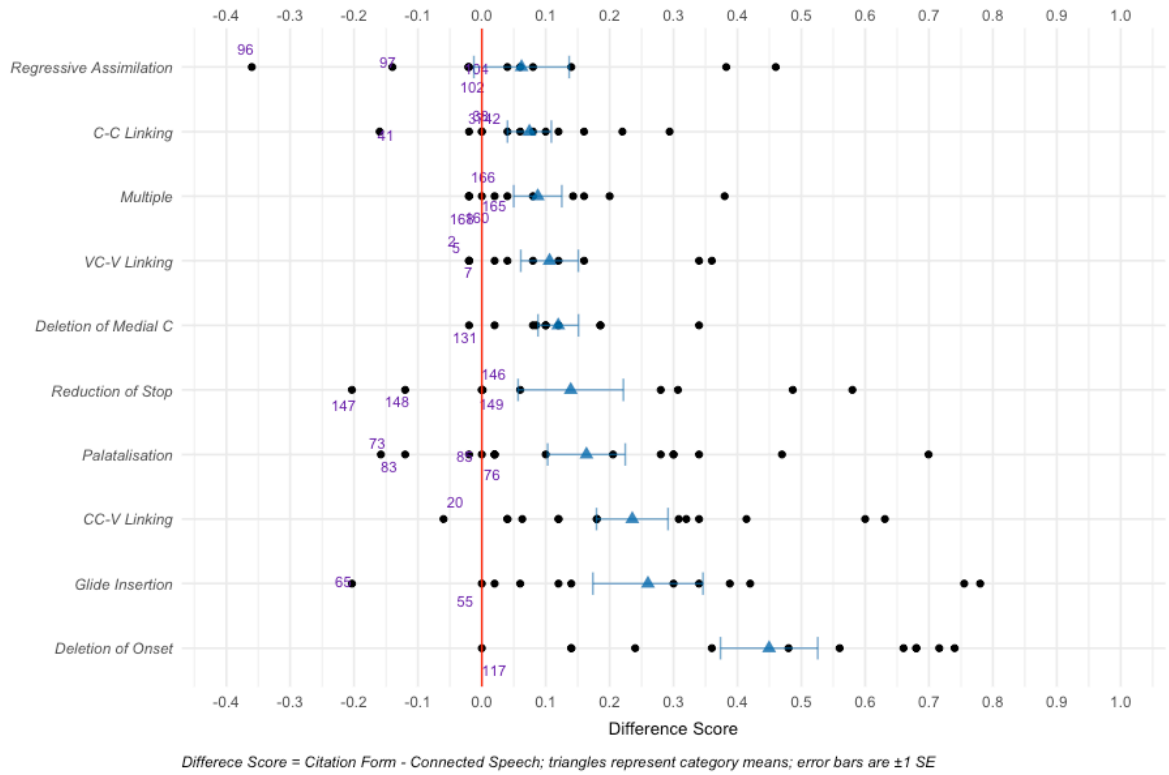
While these category-level findings were straightforward and informative, variation at the item level may have been masked – i.e. not all items within individual categories

necessarily posed equal difficulty. Certain items within a given category may have received relatively high accuracy in connected speech, potentially even surpassing that of citation form (as will be addressed in the following paragraph). In such instances, the mean accuracy of that category may have been inflated, possibly obscuring the difficulties associated with other items and the category as a whole. Hence, a closer examination of the scores for each individual stimulus item was imperative to obtain a more accurate depiction of the data.

In pursuit of this aim, a 'difference score' was computed for each item by subtracting the connected speech accuracy from the citation form accuracy. These difference scores for individual stimulus items are plotted in Figure 4.7, with category means shown as triangles. Items aligning with the vertical red line had a difference score of 0 (i.e. items transcribed with equal accuracy in connected speech and citation form). Items positioned to the left of the red line had negative difference scores (i.e. items with lower accuracy in citation form than in connected speech). Since these two subsets of items (represented by the purple serial numbers in Figure 4.7, totalling n = 28) did not present heightened processing difficulties in connected speech for the participants, they were excluded from subsequent analysis. It can be seen, however, that the large majority of items were positioned to the right of the red line, meaning they were recognised less accurately in connected speech than in citation form.

Figure 4.7

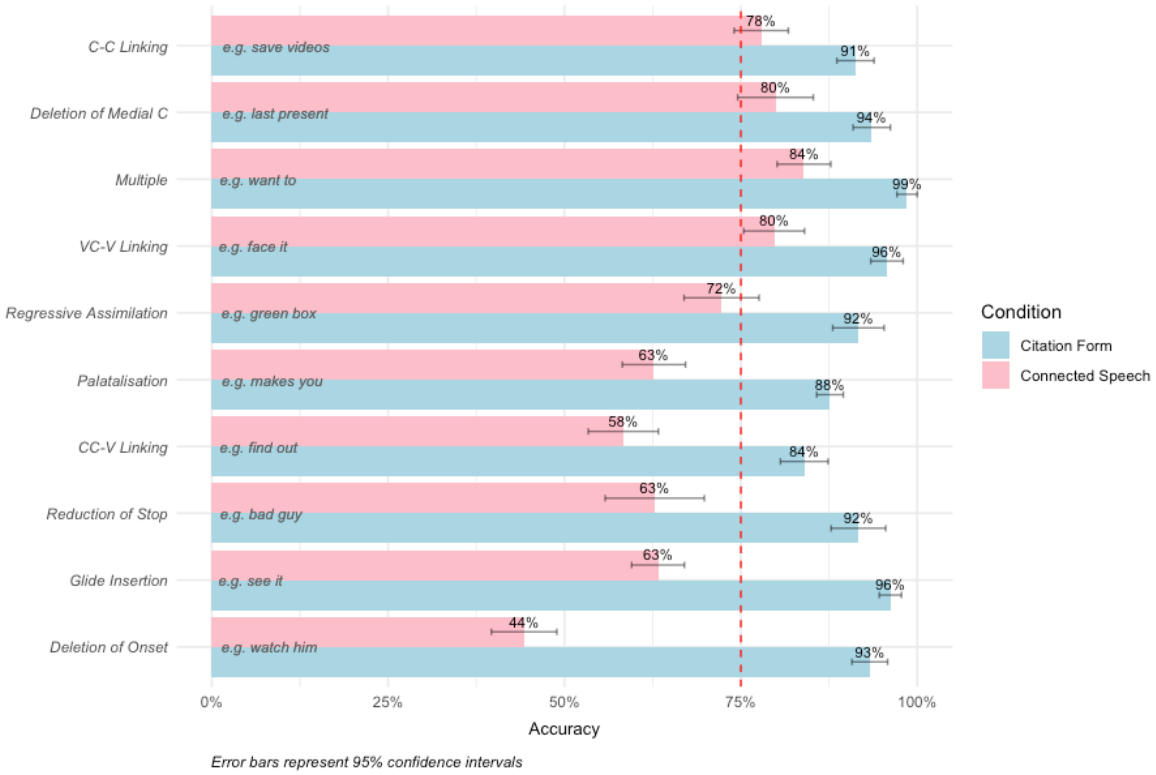
Mean Difference Scores for Individual Stimulus Items



After excluding the 28 items, a focused dataset comprising only items that showed reduced recognition accuracy in connected speech was developed. With this focused dataset, the category-level means were recalculated. As illustrated in Figure 4.8, the ranking of the most challenging categories (the bottom five) remained largely the same as in Figure 4.6 albeit with greater disparities between connected speech and citation form and slight alterations in order (cf. Figure 4.6). Notably, *Regressive Assimilation*, a category initially ranked as least problematic (with a disparity of 6 percentage points), now showed a disparity of 20 percentage points (72% accuracy in connected speech and 92% in citation form), indicating it had been previously masked by certain easier items.

Figure 4.8

Focused Dataset: Connected Speech versus Citation Form Accuracy by CSP Category



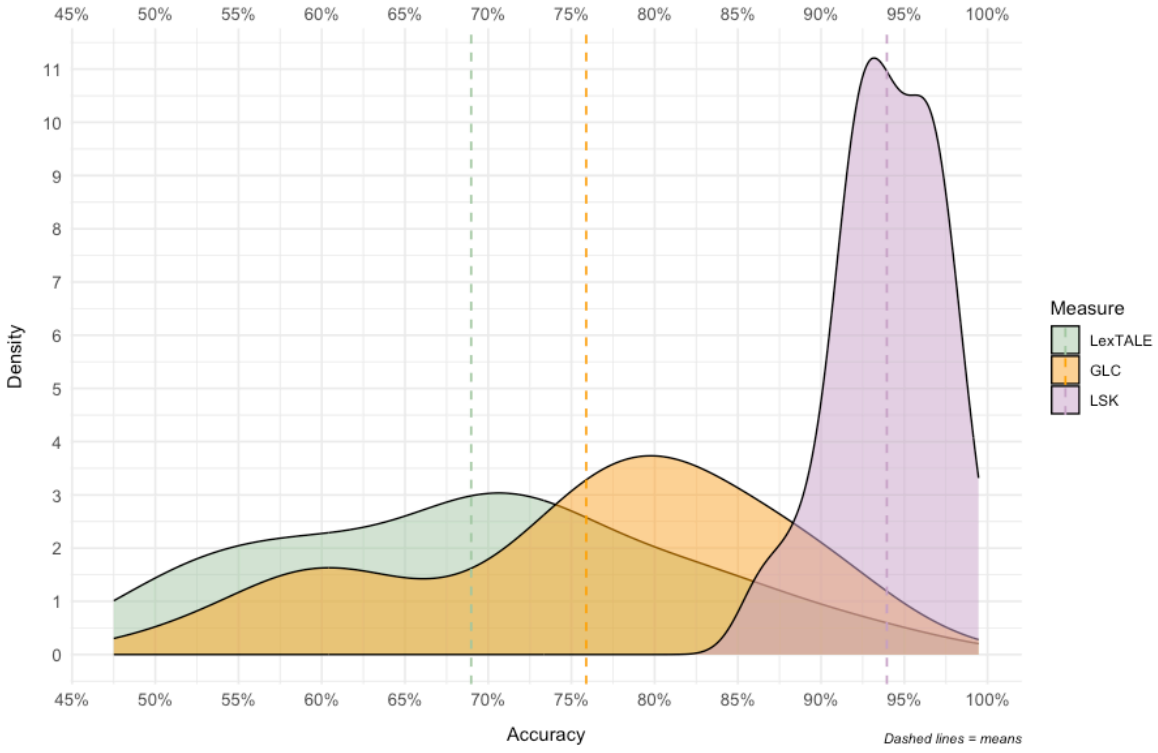
Upon examination of the focused dataset, two thresholds were established as a heuristic to identify CSP categories that were particularly challenging for the L2 learners: (i) a minimum disparity of 20 percentage points between connected speech accuracy and citation form accuracy; and (ii) 75% accuracy or lower in connected speech. In principle, both thresholds had to be met for a CSP category to be included in the Phase 2 interventions. In other words, categories with either a disparity of *less* than 20% or an accuracy in connected speech *over* 75% were considered less problematic and hence excluded from the interventions. Consequently, the bottom six categories in Figure 4.8 – *Deletion of Onset, Glide Insertion*, *Reduction of Stop*, *CC-V Linking*, *Palatalisation* and *Regressive Assimilation* (listed in descending order of accuracy disparity) – were identified as targets warranting pedagogical intervention.

4.2.3 Distributions of Supplementary Measures

In addition to the principal measures, three supplementary measures – LexTALE, GLC, and LSK – were employed to evaluate L2 participants' overall English proficiency, general listening comprehension, and L1 lexical-semantic knowledge of the target words, respectively. Figure 4.9 displays the distributions of accuracy scores across these three measures – LexTALE (green), GLC (yellow), LSK (purple) – with dashed vertical lines indicating their respective means.

Figure 4.9

Accuracy Distributions of Supplementary Measures



As anticipated, participants exhibited notably high performance on the LSK, achieving a mean accuracy of 94% (SD = 3 percentage points). This high level of performance is consistent with their advanced proficiency in English and suggests substantial familiarity

with the majority of the target words. On the other hand, while the GLC exhibited greater variability (SD = 12 percentage points), a mean accuracy of 76% reached a band sore of 7.0 in the IELTS Listening Module, which aligned with the minimum overall band score required for classification as a 'proficient' user of English. The LexTALE also showed a relatively wide distribution (SD = 11 percentage points), with a mean accuracy of 69%. This mean accuracy was lower than the 80% benchmark proposed by Lemhöfer and Broersma's (2012) for proficient users. This discrepancy was speculated to be influenced by participant's L1 backgrounds, particularly considering Lemhöfer and Broersma's data was based on Dutch participants, whose L1 shares a similar alphabetic writing system and many cognates with English. Therefore, their participants likely experienced greater ease in processing English letter strings within the stringent time constraint of this measure (a two-second timeframe for each item; cf. Section 3.3.2: *Lexical Test for Advanced Learners of English* (LexTALE)).

4.2.4 Correlations among Linguistic Measures

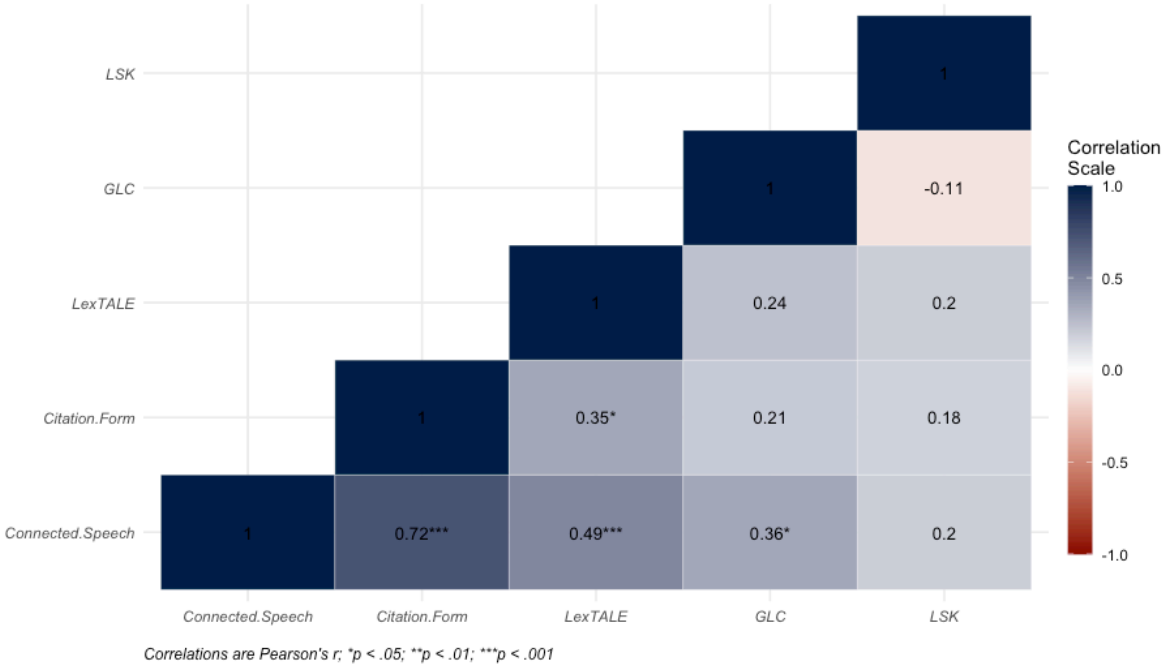
To gain a more comprehensive view of participants' performance across the five measures employed in this study, a correlation analysis was conducted to examine their interrelationships.

Figure 4.10 presents a correlation matrix for the five linguistic measures, with the significance levels indicated by asterisks. Among L2 participants, a strong correlation was observed between the connected speech and citation form dictation measures ($r = .72$, $p < .001$), underscoring that the ability to recognise words affected by CSPs is closely related to the ability to recognise the same words produced in their citation form – despite the discrepancies in accuracy rates between the two measures (cf. Sections 4.2.1 and 4.2.2

above). The connected speech dictation also showed moderate correlations with both the LexTALE (r = .49, p < .001) and GLC (r = .36, p < .05), suggesting that the ability to recognise CSP-affected words is associated, to some extent, with overall English proficiency and general listening comprehension.

Figure 4.10

Correlation Matrix for Linguistic Measures: Phase 1 Diagnosis



Similarly, the citation form dictation correlated moderately with the LexTALE (r = .35, p < .05), supporting the association between recognising words in their citation form and overall proficiency. However, both the LexTALE and citation form dictation demonstrated only weak and non-significant correlations with the GLC (r = .24 and .21, respectively, p > .05). Therefore, in this data, there was no evidence showing that overall proficiency in English (as measured by the LexTALE) and the ability to recognise words in their citation form could be associated with the ability to comprehend extended spoken

discourses.

The weak and non-significant correlations between the LSK and all the other measures were expected due to the deliberate control over vocabulary level – over of 70% of the target words were of beginner level, and none were of advanced level (cf. Table 3.2). Consequently, participants with advanced proficiency, like those in the current L2 group, were able to recognise the orthographic form and meaning of the majority of these words (mean accuracy = 94%, as shown in Figure 4.9), despite greater variability in other linguistic measures.

4.3 Summary of Phase 1 Study

In Phase 1, ten CSP categories were selected for investigation based on five criteria: inter-word environment, non-lexicalisation, universality, high frequency, and high saliency. A stimulus inventory was developed, comprising 116 experimental stimuli along with 58 distractors. The experimental stimuli included a balanced range of relevant phonological contexts (target segments involved) for each CSP category and underwent two pilot experiments (assessing grammaticality and predictability). Following this, the items were audio recorded, and a phonetic analysis was conducted to ensure a clear presence of target CSPs.

Two principal measures (connected speech dictation and citation form dictation) were employed based on the devised stimuli. Additionally, three supplementary measures (LexTALE, GLC, and LSK) were administered to assess participants' overall proficiency, L2 general listening comprehension, and L1 lexical-semantic knowledge, respectively. All measures were programmed into a single online experiment via Gorilla and completed

by 50 adult Mandarin-speaking learners of L2 English. As a comparison group, 20 adult L1 (native English) participants completed the two dictation measures to establish baseline performance.

The results indicated that the L2 group experienced significantly greater difficulty in processing connected speech than citation form, compared to the L1 group. Additionally, the degree of difficulty varied across CSP categories, with disparities in accuracy between connected speech and citation form ranging from 13 to 49 percentage points. Critically, six (out of ten) CSP categories were identified as particularly challenging, each characterised by: (i) a minimum disparity of 20 percentage points in accuracy between connected speech and citation form; and (ii) 75% accuracy or lower in connected speech. These categories were *Deletion of Onset, Glide Insertion*, *Reduction of Stop*, *CC-V Linking*, *Palatalisation*, and *Regressive Assimilation*.

Regarding the supplementary measures, the LSK exhibited a high mean accuracy (94%), as expected. On the other hand, the LexTALE yielded a lower mean accuracy (69%) compared to the original study by Lemhöfer and Broersma (2012), likely reflecting differences in participants' L1 backgrounds. The mean accuracy on the GLC (76%) met the minimum threshold associated with proficient users in the IELTS Listening. Correlation analysis indicated that participants' performance on connected speech dictation, citation form dictation, and the LexTALE were positively and significantly correlated. However, general listening comprehension only showed a significant correlation with connected speech dictation. As expected, the LSK did not significantly correlate with any other measure, reflecting the controlled vocabulary level and participants' overall high familiarity with the target words.

The following Phase 2 study built on these findings by further examining the effectiveness of pedagogical interventions targeting the six most challenging CSP categories (as identified above). It focused particularly on the extent to which different pedagogical approaches may facilitate the intelligibility and comprehensibility of connected speech for L2 learners, as well as their relationships with supplementary linguistic measures before and after the interventions.

Chapter 5: Phase 2 Study – Intervention

The Phase 2 study was designed to investigate the effectiveness of different pedagogical interventions targeting the particularly challenging CSP categories identified in the Phase 1 study. This phase involved developing distinct intervention programmes and administering a pre-test, interventions, and a post-test to adult Mandarin-speaking learners of L2 English. Similar to Phase 1, this chapter is structured into methods, results, and summary sections.

5.1 Methods

5.1.1 Participants

Building upon the findings of Phase 1, the Phase 2 study recruited participants with similar language backgrounds, specifically focusing on adult Mandarin-speaking learners of L2 English who were university students in Taiwan. Participants either were currently enrolled in, or had previously completed, degree programmes in the English language, foreign languages and literatures, (applied) linguistics, TESOL, or related disciplines. This learner demographic was chosen for two primary reasons: (1) given their academic foci, they were considered a motivated population who were particularly relevant beneficiaries of the interventions aimed at enhancing L2 English speech perception; (2) they were expected to possess generally higher proficiency in English compared to students from non-language-related disciplines. The minimum proficiency requirement for participation in this study was established at the CEFR B2 level (upper-intermediate; cf. Appendix B), in recognition that many undergraduate students with language-related majors within the Taiwanese EFL context have yet to attain the CEFR C1 level (advanced).

Through email correspondences, thirteen collaborating universities in Taiwan

(comprising 11 national and 2 private universities) were identified. At these universities, relevant departments or faculty members agreed to disseminate the information about the study by sharing a digital recruitment poster with their students. Students who were interested in participating registered via an online survey link, where they were required to (1) confirm that they were currently enrolled in or had previously studied one of the relevant disciplines; and (2) provide a preferred email address through which they could receive a participant ID code, instructions, and access to the study URLs. Throughout the study, all correspondence with participants used their registration email. At no point were participants asked to disclose their personal names or the names of their academic institutions; they were referred to only by their unique participant ID code.

Upon registration, participants were randomly assigned to one of three intervention conditions (described in detail in Section 5.1.4). The initial recruitment target was a minimum of 20 participants per intervention condition. However, due to the randomisation process and varying completion rates across groups, there was an imbalance in group composition. Specifically, the Experimental Group 1 initially had a higher number of participants (n = 9) reporting an English proficiency level below CEFR B2, compared to the other groups. As shown in Table 5.1, participants reporting proficiency below B2 and those who were unable to provide an equivalent standardised proficiency certificate were excluded. For participants with unverified proficiency, similar inclusion and exclusion criteria as in Phase 1 were applied (cf. Table 4.1). This entailed calculating the mean scores and standard deviations for the LexTALE, GLC, and LSK measures based on data from participants with certified proficiency at B2 or above. Cut-off thresholds for each measure were set at two standard deviations below the mean, approximating the lower limit of the 95% of a normal distribution (though not assuming

the data followed the normal distribution perfectly). Following these procedures, one additional participant from the Experimental Group 2 and one from the Control Group were retained after proficiency verification. The final numbers of participants included in the analyses were 20 for the Experimental Group 1, 23 for the Experimental Group 2, and 23 for the Control Group.

Table 5.1

Summary of Participant Recruitment and Final Inclusion by Group

	Experimental Group 1	**Experimental Group 2**	**Control Group**	**Total**
Completed Study	31	27	25	83
Excluded: Proficiency Below B2	– 9	– 4	– 2	– 15
Excluded: Proficiency Unverified	– 2	– 1	– 1	– 4
Included Upon Proficiency Verification	+ 0	+ 1	+ 1	+ 2
Final Inclusion	**20**	**23**	**23**	**66**

5.1.2 Target CSP Categories and Stimuli

The six CSP categories identified as particularly challenging in Phase 1 were *Deletion of Onset, Glide Insertion*, *Reduction of Stop*, *CC-V Linking*, *Palatalisation*, and *Regressive Assimilation*. The same stimulus inventory described in Section 4.1.3 was used to examine these categories (cf. Appendix C for the complete list of stimulus items), with the exclusion of items that Phase 1 results indicated did not present heightened difficulties in connected speech (cf. Section 4.2.2).

As summarised in Table 5.2, following the exclusion of 16 items, a total of 57 items from

the original stimulus inventory were retained, covering the six target CSP categories. These items were divided into two sets, for the pre-test and post-test, respectively. It is important to note that none of these stimulus items were included in the instructional content during the interventions. This ensured that words and phrases featuring target CSPs did not appear as examples or practice materials, thereby minimising the risk of practice effects influencing post-test performance.

Table 5.2

Stimulus Items Used in Phase 2 Study

	Original Items	**Excluded Items**	**Retained Items**
Regressive Assimilation	10	4	6
Palatalisation	15	4	11
CC-V Linking	14	1	13
Reduction of Stop	10	4	6
Glide Insertion	12	2	10
Deletion of Onset	12	1	11
Total	**73**	**16**	**57**

5.1.3 Implementation Procedures

As previously mentioned, the implementation procedures for Phase 2 comprised three sequential components: a pre-test, pedagogical interventions, and a post-test. All participants completed the pre-test and post-test. For the interventions, participants were randomly assigned to one of three conditions, each representing a distinct pedagogical approach – rule-based, noticing-based, or implicit (serving as a control group). To mitigate potential stimulus-related biases, participants within each group were further

subdivided to counterbalance the stimulus sets presented at the pre-test and post-test. Specifically, half of the participants encountered Set 1 at the pre-test and Set 2 at the post-test, while the other half experienced the reverse order.

Pre-test Procedure

In the pre-test, participants completed a demographic questionnaire, followed by six linguistic measures (cf. Section 3.3) administered via Gorilla. The sequence of measures was as follows: LexTALE, connected speech dictation along with comprehensibility rating, L2 general listening comprehension (GLC), citation form dictation, and L1 lexical-semantic knowledge test (LSK). This specific order was designed to prevent participants from gaining any advantage from familiarity with the target words, as explained in Phase 1 (cf. Section 4.1.4). For instance, administering the citation form dictation prior to the connected speech dictation and comprehensibility rating, or the LSK before the principal measures, could have compromised the validity of the results by priming participants. Participants were required to complete the pre-test within three days of receiving their unique ID code (used for logging into the experiment) and the corresponding study URL.

Intervention Procedure

Following the pre-test, participants were provided access to their assigned online pedagogical programmes, which consisted of one introductory and three instructional sessions. Each instructional session was approximately 20 minutes in length. Completion of all the intervention sessions, along with an engagement check questionnaire, was mandatory before progressing to the post-test. The engagement check questionnaire required participants to select sentences they had practised during the interventions from a list of twelve items – six of which were actual sentences (i.e. those practised in the

intervention sessions) and six were fillers. This procedure was intended to ensure participants had actively engaged with the instructional materials.

Additionally, to minimise recency effects, wherein the most recently presented materials tend to be more readily retained, participants were required to complete the instructional sessions in predetermined sequences. Specifically, the six CSP categories were divided into three instructional sessions. Therefore, as shown in Table 5.3, there were six possible sequences (i.e. all possible permutations), which were assigned approximately evenly to participants to ensure counterbalanced sequential exposure to instructional materials. Participants were instructed to complete all intervention components, including the instructional sessions and the engagement check questionnaire, within seven days of receiving the materials.

Table 5.3

Counterbalanced Sequences of Instructional Sessions

Sequence	**Session 1**	**Session 2**	**Session 3**
1	*CC-V Linking / Glide Insertion*	*Reduction of Stop / Deletion of Onset*	*Palatalisation / Regressive Assimilation*
2	*CC-V Linking / Glide Insertion*	*Palatalisation / Regressive Assimilation*	*Reduction of Stop / Deletion of Onset*
3	*Reduction of Stop / Deletion of Onset*	*CC-V Linking / Glide Insertion*	*Palatalisation / Regressive Assimilation*
4	*Reduction of Stop / Deletion of Onset*	*Palatalisation / Regressive Assimilation*	*CC-V Linking / Glide Insertion*
5	*Palatalisation /*	*CC-V Linking /*	*Reduction of Stop /*

	Regressive Assimilation	*Glide Insertion*	*Deletion of Onset*
6	*Palatalisation / Regressive Assimilation*	*Reduction of Stop / Deletion of Onset*	*CC-V Linking / Glide Insertion*

Post-test Procedure

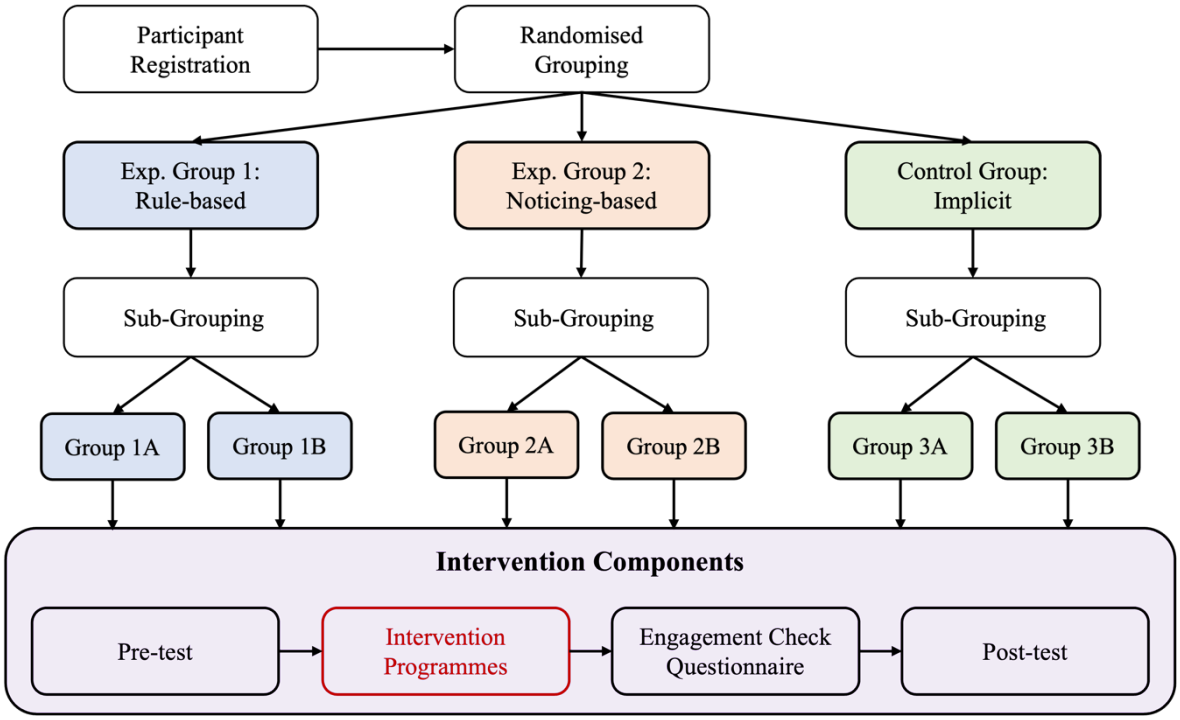
In the post-test, participants completed the same battery of linguistic measures as in the pre-test, albeit using a different set of stimuli. The sequence of measures remained identical to that of the pre-test. Participants were required to complete the post-test within three days of receiving the study URL. Upon completion of the entire intervention process – including the pre-test, intervention programme, and post-test – participants received a digital coupon worth NTD$300 [29] (approximately £7.30) as a modest token of appreciation. Additionally, participants were provided a personalised performance report following the completion of the study.

Figure 5.1 below provides an overview of the intervention study procedures. These structured procedures were designed to ensure the reliability and validity of the data collected, facilitating the evaluation of the effectiveness of different pedagogical approaches for teaching CSPs.

Figure 5.1

[29]This amount was roughly equivalent to 1.5 – 2 hours of minimum wage in Taiwan at the time of the study (2023 – 2024), based on data from the Ministry of Labor, Executive Yuan, Taiwan.

Overview of Phase 2 Intervention Procedures



Note. Sub-grouping within each intervention condition was implemented to counterbalance the stimulus sets used across the pre-test and post-test.

5.1.4 Pedagogical Approaches

As outlined above, participants were randomised into one of the three intervention conditions, each representing a distinct pedagogical approach: a rule-based condition, a noticing-based condition, and an implicit condition. The basis for comparing these approaches lay in their theoretical distinctions and contrasting instructional designs in the current study.

Theoretical Distinctions

1. **Rule-based Condition:**

   This condition is grounded in deductive, explicit instruction, wherein learners are provided with direct rule explanations of the target forms, typically accompanied by

the use of metalanguage (Norris & Ortega, 2000). The primary aim is to promote leaners' conscious understanding of the target linguistic structures (i.e. CSP mechanisms in the current study), thereby facilitating the application of these rules when processing novel items.

2. Noticing-based Condition:

This condition draws on the Noticing Hypothesis (Schmidt, 1990), which posits that input must be consciously noticed to become intake, although such noticing does not necessarily require deliberate effort if the learning tasks are designed to focus attention on the target forms. Aligned with inductive learning, this approach exposes learners to targeted examples and guides them to attend to specific language features without providing explicit rule explanations. Learners are expected to infer generalisations or patterns through guided instruction (Decoo, 1996) to support their processing of novel items.

3. Implicit Condition:

This condition reflects implicit instruction, wherein the target forms are embedded within meaning-focused communicative discourse, with learners' attention directed towards comprehending message meaning rather than specific linguistic forms. This approach is predicated on the assumption that learners may incidentally acquire the target forms through input exposure without awareness (Ellis, 2009), as no explicit instruction or guided attention is provided. This condition served as a control group in the current study, with the instruction materials designed to include the same CSP samples as for the experimental conditions (see *Control Group: Implicit Condition* in *Instructional Designs* below for implementation details).

Instructional Designs

The instructional designs of the three intervention conditions carefully aligned with their respective theoretical underpinnings. The rule-based group received deductive, explicit instruction, with direct metalinguistic explanations detailing the occurrences and mechanisms of each CSP category. The instructional focus was placed on developing learners' understanding of the rules governing the usage of different CSP categories, enabling them to apply these rules when encountering other lexical items similarly affected by the same CSPs.

The noticing-based group engaged in inductive, guided instruction, designed to facilitate the noticing of differences between dictionary pronunciation (i.e. the citation form) and real-life pronunciation (i.e. connected speech). No explicit rule explanations were provided; instead, the instructional objective was to encourage learners to discern CSP patterns through comparing these forms and focusing their attention on pronunciation differences in targeted examples.

In contrast, the implicit (control) group was exposed to extended meaning-focused spoken discourses, during which learners completed listening comprehension tasks without any direct focus on pronunciation or attention to CSP features. This condition was designed to provide incidental exposure to the target CSPs, while learners' attention remained on comprehending the message meaning. The key characteristics of the three pedagogical approaches are summarised in Table 5.4.

Table 5.4

Summary of Pedagogical Approaches for the Intervention Study

	Exp. Group 1: Rule-Based Condition	**Exp. Group 2: Noticing-Based Condition**	**Control Group: Implicit Condition**
Direct Focus on CSPs	Yes	Yes	No
Instructional Method	Deductive, explicit rule explanations	Inductive, guided attention without rule explanations	Implicit exposure to meaning-focused texts
Instructional Objective	Understand rules underlying different CSP categories	Notice differences in dictionary vs. real-life pronunciation	Check listening comprehension
Instructional Materials	Short video clips: individual sentence	Short video clips: individual sentences	Longer video clips: extended discourses (containing the same sentences within them as experimental groups)
Practice Activities	Identify CSP mechanisms in targeted examples; hear them in connected speech	Compare pronunciation changes in targeted examples; hear them in connected speech	Take notes while listening to passages; answer comprehension questions

All groups received highly controlled input with identical sample words and sentences featuring the target CSPs, despite differences in instructional methods and practice activities. The two experimental groups practised with short video clips containing individual sentences, focusing on CSP-affected phrases within sentential contexts. In contrast, the control group was exposed to longer video clips (each featuring an approximately one-minute discourse) that included the same sentences presented to the

experimental groups, thereby incorporating the target CSPs implicitly into their listening comprehension practice. Further, one of the comprehension questions for each longer video clip was strategically designed to elicit specific information from the exact sentence containing the target CSPs (i.e. the sentence practised in the experimental groups). This ensured that the control group consistently had the opportunity to encounter the same CSP examples, albeit implicitly, and did not disregard these inputs due to their irrelevance to completing the listening comprehension task.

Throughout the interventions, all groups encountered a total of 30 video clips (5 for each CSP category) featuring UK English. These materials were curated through a targeted search of CSP-prone phrases on YouGlish[30], an online tool that retrieves YouTube video contents containing specified words or phrases. The material selection criteria included a clear presence of the target CSPs, appropriate content level (considering speech rate, topic and vocabulary familiarity, and speaker accent), and diversity of sources (including UK parliament, research organisations, museums, educational platforms, TED Talks, independent media). All materials were either licensed under Creative Commons[31] or used with direct permission from the content creators. These materials were considered authentic, in the sense that they were originally created for purposes other than language teaching, thereby representing natural, dynamic, real-life connected speech, particularly in contrast to carefully articulated pronunciation typically found in many foreign

[30]YouGlish (https://youglish.com/) is an online search tool that enables users to locate videos on YouTube containing specific words or phrases. It pinpoints the exact sentences in which the searched terms appear and allows users to select among different language varieties, such as US, UK, or Australian English.

[31] Creative Commons (https://creativecommons.org/) is an international organisation that provides standardised licences, enabling content creators to protect their works while allowing others to legally copy, share, distribute, and use the content under specified conditions.

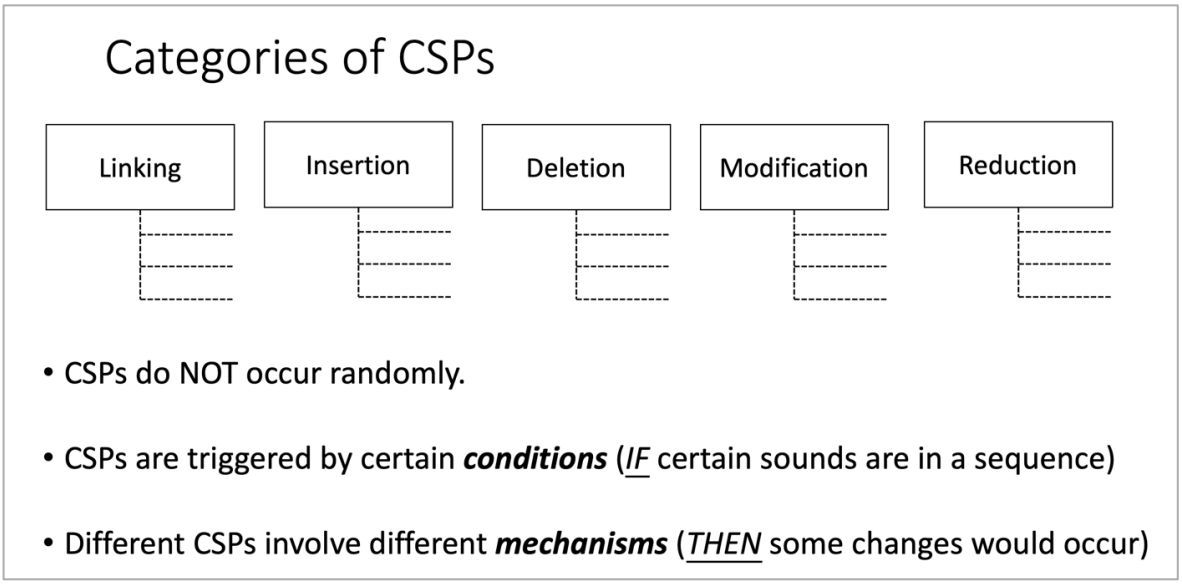
language teaching-oriented materials.

The following sections provide illustrative examples of the instructional content and practice activities for each group, accompanied by sample slides used during the instruction. A comprehensive set of instructional materials is included in Appendix E .

Experimental Group 1: Rule-Based Condition. In the rule-based condition, participants received explicit metalinguistic instruction in CSP rules. During the introductory session, participants were presented with a definition of CSPs and were informed that they would learn the rules of various CSP categories systematically through a 'conditional structure'. This structure involved the explanation of (1) the conditions (the 'if' part) that trigger the occurrence of the CSP, referring to the specific phonological contexts of individual CSP categories; and (2) the mechanisms (the 'then' part) that result from these phonological conditions, indicating the changes in pronunciation. Figure 5.2 presents a slide from the introductory session that includes the description of the conditional structure employed throughout the subsequent instructional sessions.

Figure 5.2

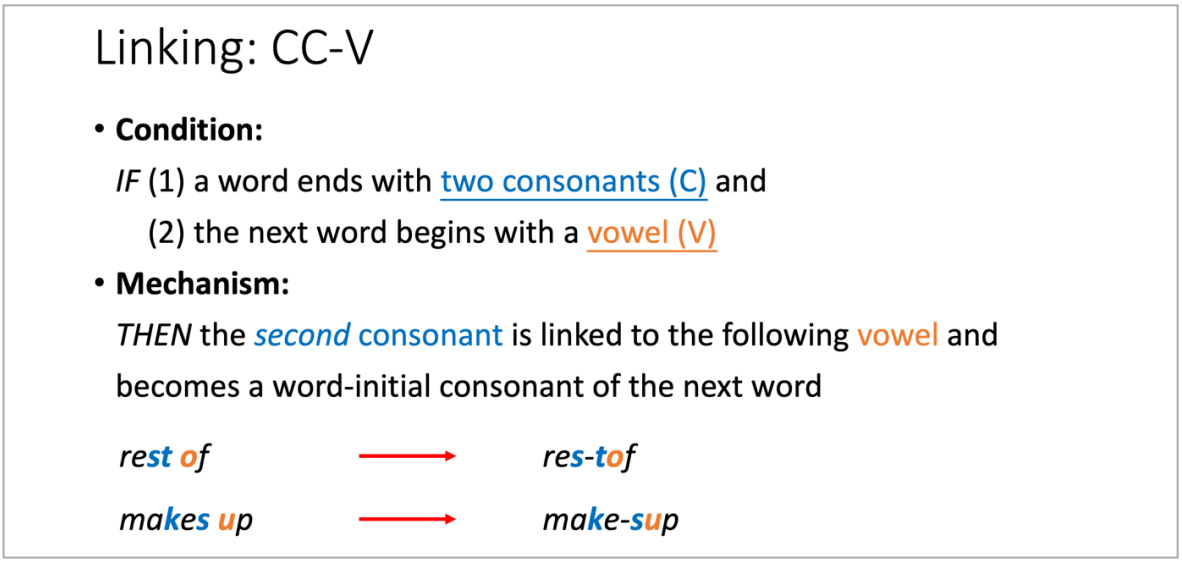
Sample Slide: Rule-based Condition – Introductory Session



Following this structure, the instruction for each CSP category began with explicit explanations of its condition and mechanism, supported by two targeted examples (Figure 5.3). Given that not all participants were familiar with the International Phonetic Alphabet (IPA), changes in pronunciation were visually presented using sound-approximating spellings, colour-coded letters, and punctuation marks (e.g. hyphen, under tie, deleatur, brackets), rather than IPA transcription.

Figure 5.3

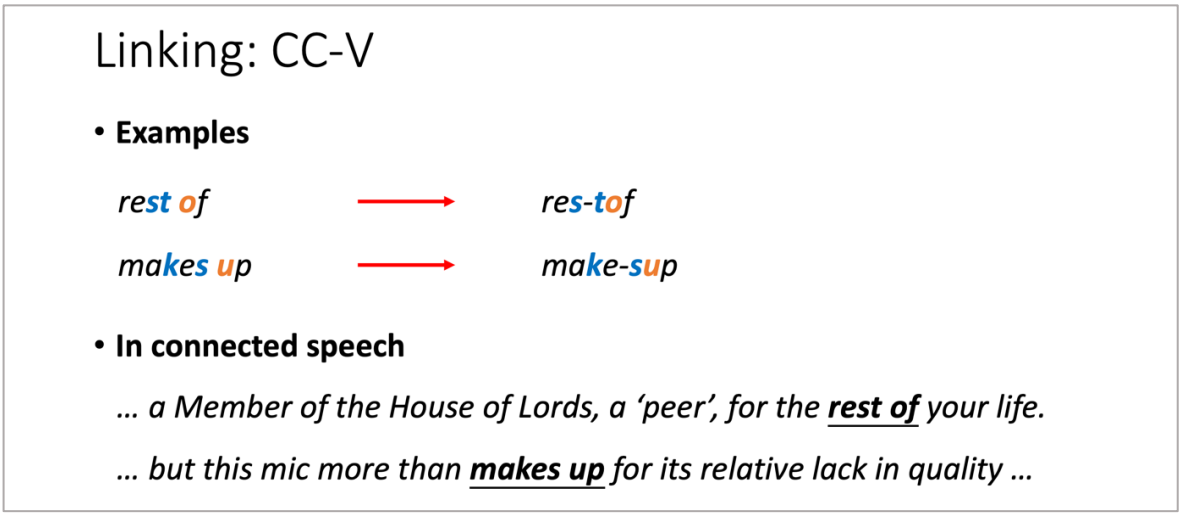
Sample Slide: Rule-based Condition – Rule Explanations with Targeted Examples



Participants were subsequently provided with connected speech contexts for the examples, and were told that in these sentential contexts, the pronunciation of the example phrases would feature the target CSPs (Figure 5.4). These sentences, as extracted from authentic online videos, may vary in length. Thus, if necessary, longer sentences were slightly edited down to ensure a focus on CSP-affected phrases (for presentation in the slide) while preserving meaningful thought groups.

Figure 5.4

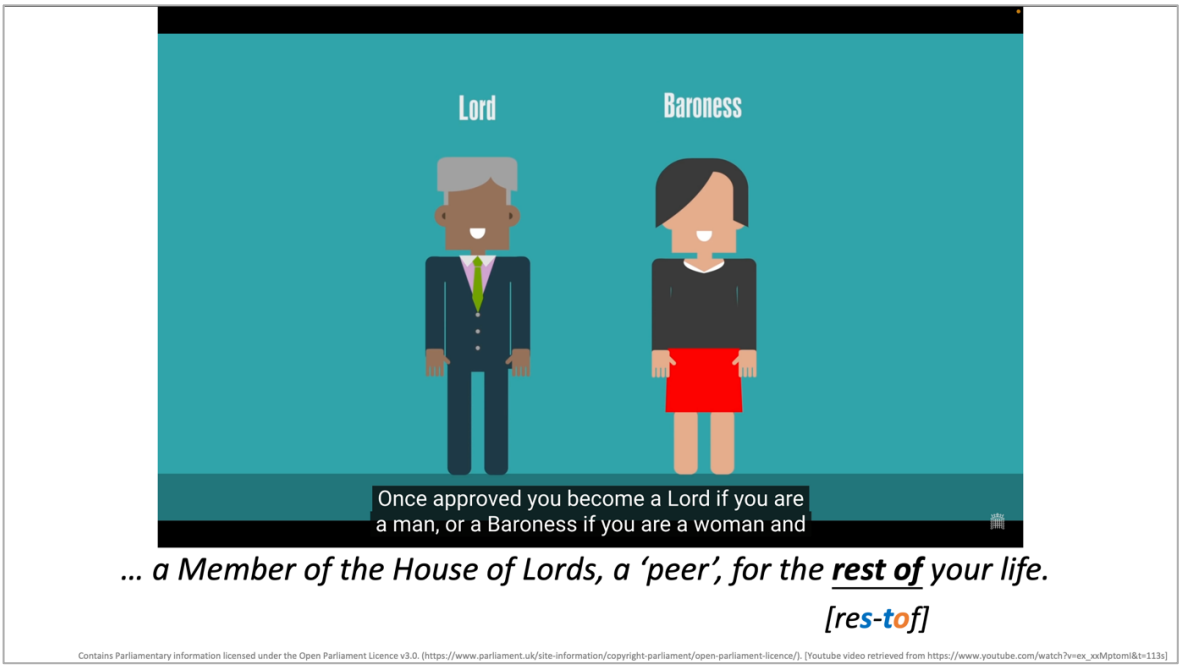
Sample Slide: Rule-based Condition – Connected Speech Contexts



After familiarisation with the contexts, participants watched and listened to short video clips of these sentences twice (Figure 5.5). They were instructed to concentrate on how the target phrases were pronounced with CSPs, as opposed to their respective citation forms. Visual representations of connected speech pronunciation were displayed beneath the example phrases to support the aural input.

Figure 5.5

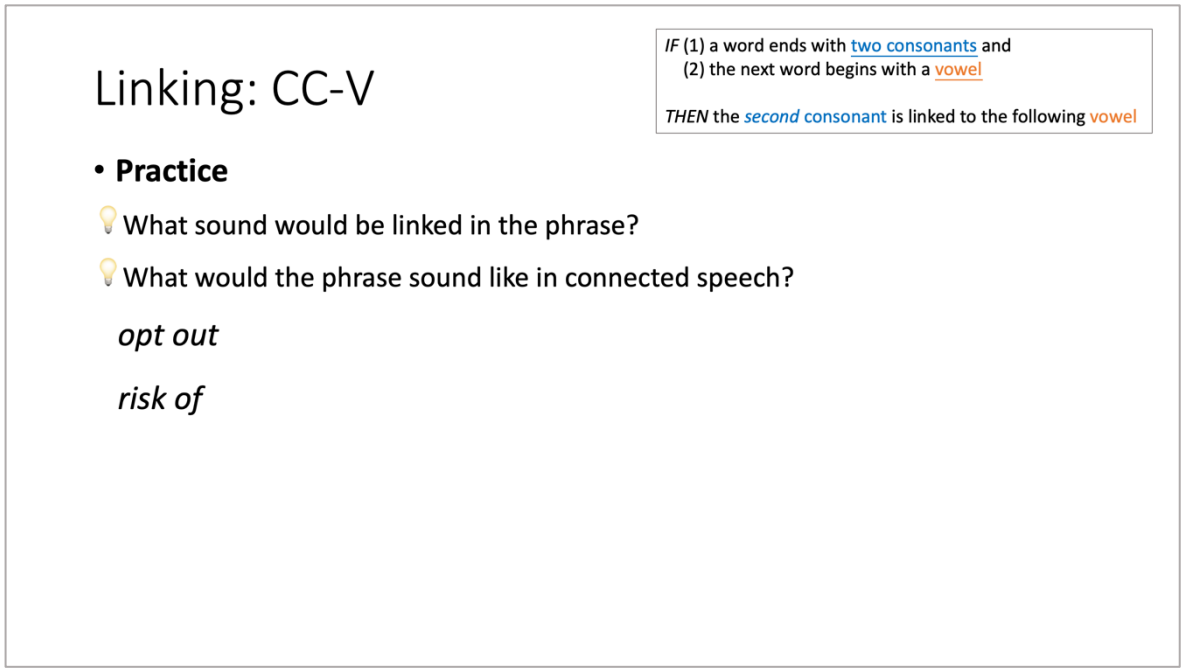
Sample Slide: Rule-based Condition – Video Clip Material for Targeted Example



Following this, the session progressed to a practice activity involving two additional phrases. Participants were instructed to identify within these phrases the sounds that would be affected by the target CSPs, and to form mental representations of their connected speech pronunciation, based on the previously introduced rules (Figure 5.6). The CSP condition and mechanism were reiterated in the slide as a reminder. Participants were given time to work on this practice and were encouraged to pause the instructional video if more time was needed.

Figure 5.6

Sample Slide: Rule-based Condition – Practice (Identification of CSP-affected Sounds)

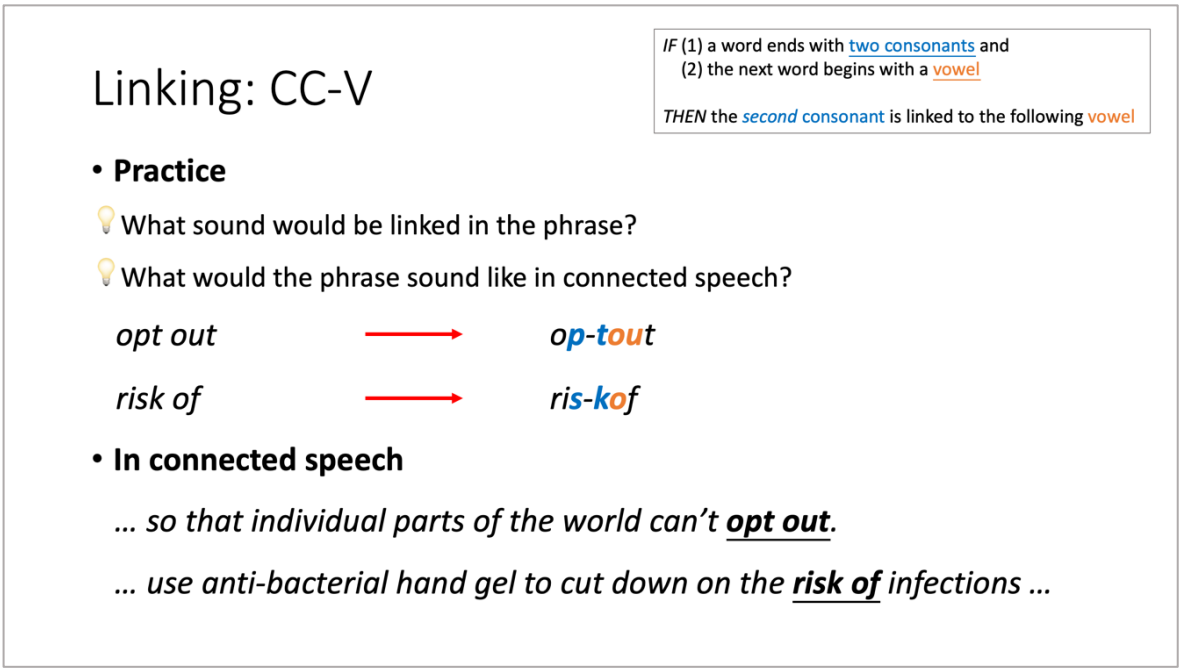


Upon the identification of CSP-affected sounds, participants were presented with connected speech contexts for these phrases (Figure 5.7). They were instructed to apply their mental representations of the connected speech pronunciation to the given sentential contexts in preparation for the subsequent video clips exemplifying these phrases.

Figure 5.7

Sample Slide: Rule-based Condition – Practice (Familiarisation with Connected Speech

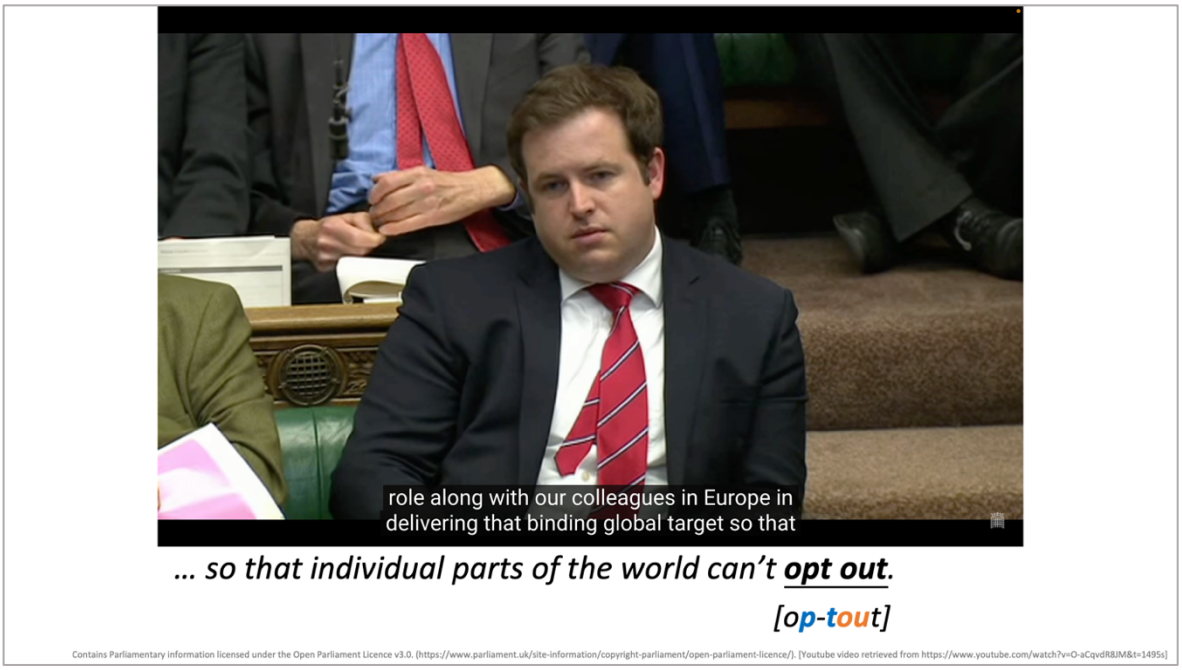
Contexts)



Participants then watched and listened to short video clips for each phrase. They were instructed to concentrate on how the target phrases were pronounced with CSPs, verifying whether their mental representations aligned with the actual connected speech pronunciation in the given contexts (Figure 5.8).

Figure 5.8

Sample Slide: Rule-based Condition – Practice (Video Clip Material)

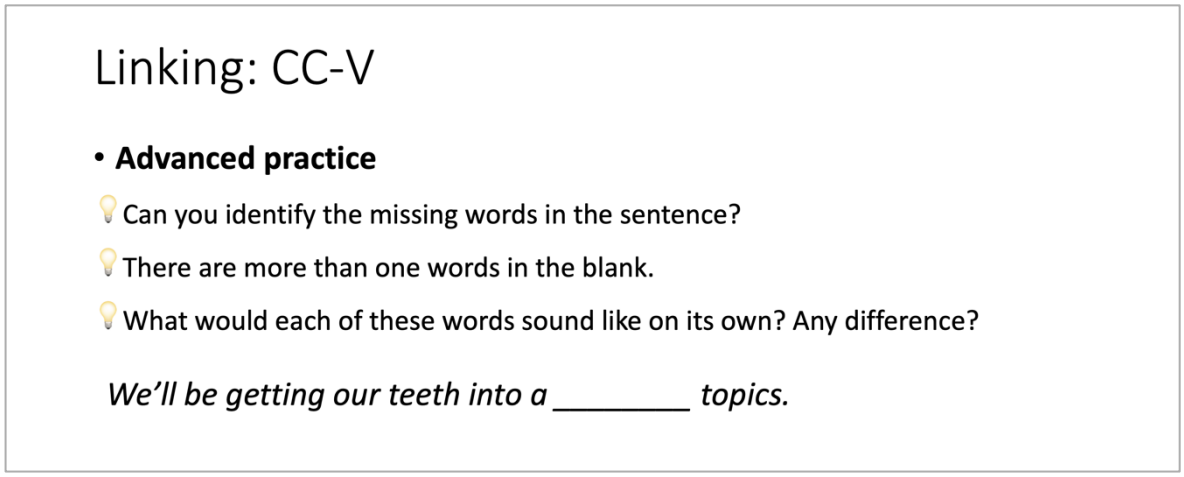


Subsequently, an advanced practice task was introduced, where participants were presented with a sentence containing two missing words affected by the target CSPs. They were instructed to identify the missing words while watching and listening to a short video clip of this sentence, and to reflect on how the pronunciation of these words (as pronounced by the speaker) differed from their respective citation forms (Figure 5.9).

Figure 5.9

Sample Slide: Rule-based Condition – Advanced Practice (Identification of Missing

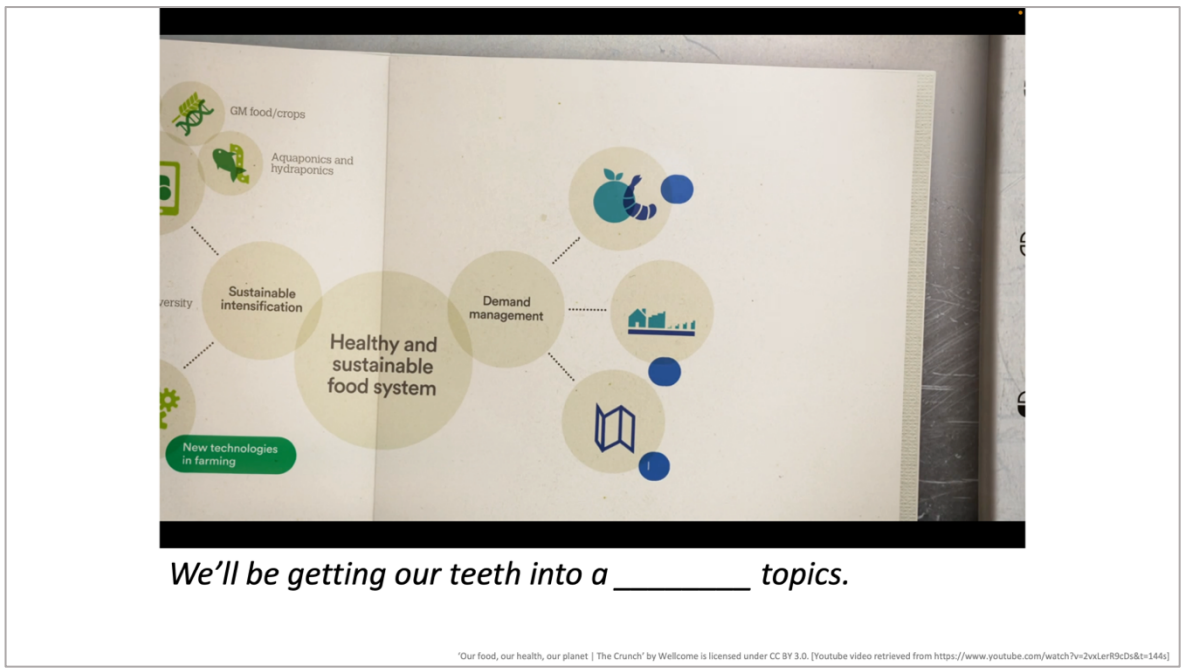
Words)



The video clip was played twice – while participants were allowed to replay it as many times as needed – before the correct answers were provided (Figure 5.10). This advanced task marked the conclusion of the instruction for each CSP category.

Figure 5.10

Sample Slide: Rule-based Condition – Advanced Practice (Video Clip Material)



Note. The intended words missing in this sentence were *range of*.

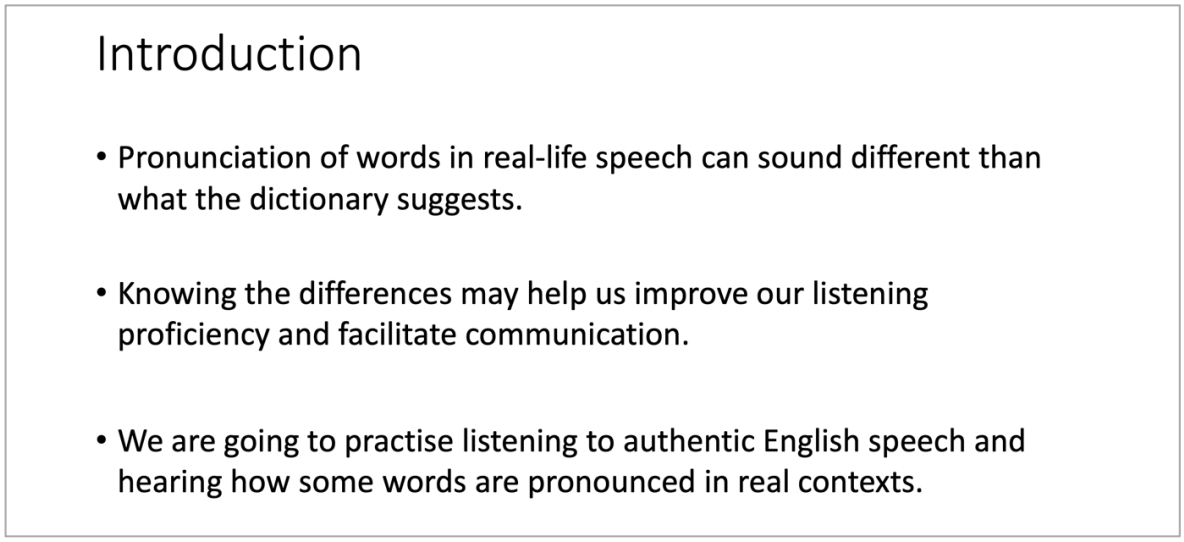
In summary, the CSP instruction for the rule-based group followed a structured progression, beginning with metalinguistic rule explanations and two targeted examples, followed by a practice activity involving the identification of CSP-affected sounds in two additional phrases, and concluding with an advanced word identification task. Throughout this process, participants were presented with five targeted examples per CSP category, all demonstrated within authentic connected speech contexts. This structure was consistently applied across all six CSP categories, ensuring tightly controlled input and instructional procedures throughout the intervention.

Experimental Group 2: Noticing-based Condition. In the noticing-based condition, participants received induction-oriented instruction designed to guide their attention to differences between dictionary pronunciation (i.e. the citation form) and real-life pronunciation of words. During the introductory session, participants were informed that the purpose of the intervention programme was to practise hearing such differences in authentic speech[32] contexts. Importantly, technical terminologies such as 'CSPs' and 'connected speech' were not introduced, nor were any definitions of CSPs provided. Figure 5.11 presents a sample slide from the introductory session outlining the focus of the intervention.

Figure 5.11

[32]To avoid the technical term *connected speech*, a more accessible alternative *authentic speech* was adopted in the instruction for the noticing-based condition. Therefore, in describing the instructional procedures for this condition, the term *authentic speech* is used.

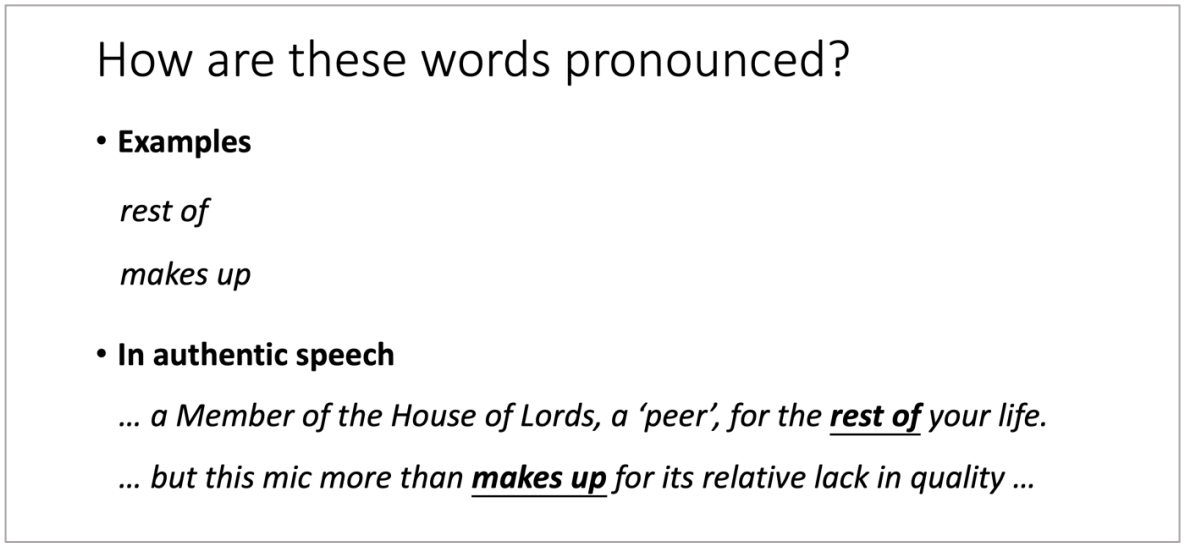
Sample Slide: Noticing-based Condition – Introductory Session



In this condition, the instructional sessions began directly with targeted example phrases. Participants were first asked how the constituent words within these phrases would be pronounced individually, and then were presented with sentential contexts containing these phrases. They were invited to consider whether and how the pronunciation of these words would change when spoken within authentic speech contexts (Figure 5.12). Note that while all the example phrases and sentences were identical to those used in the rule-based condition, participants in the noticing-based condition were *not* introduced to any CSP category labels (e.g. *CC-V Linking*) or phonological mechanisms (e.g. a word-final consonant linked to the following word-initial vowel).

Figure 5.12

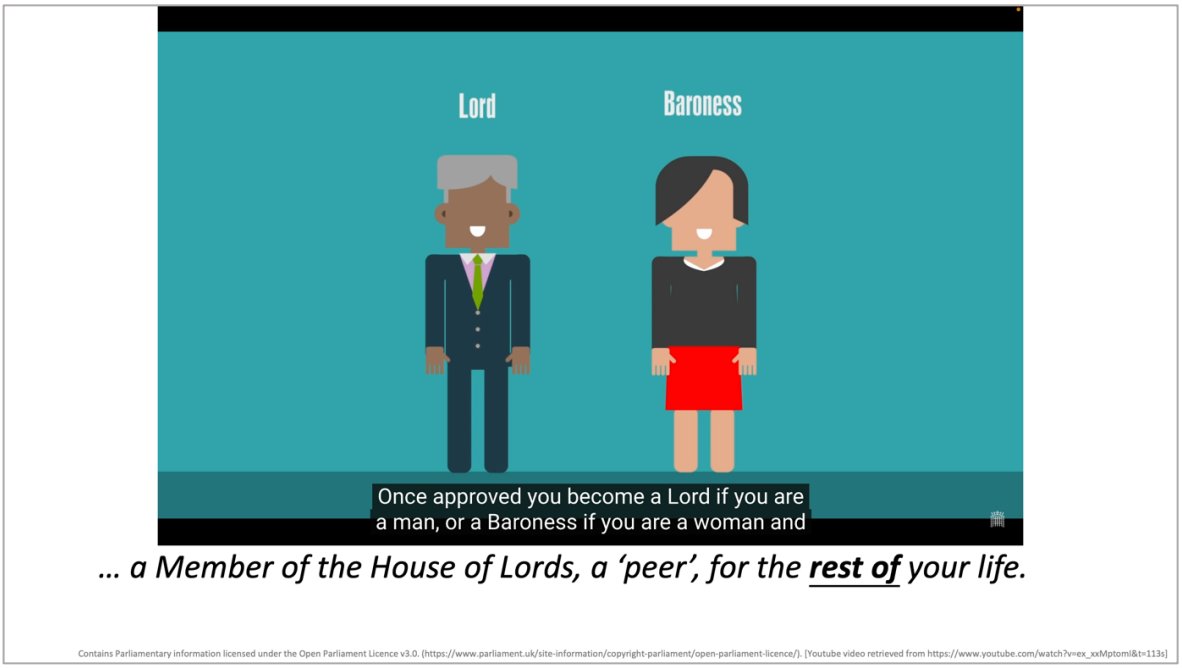
Sample Slide: Noticing-based Condition – Guided Instruction with Targeted Examples



Participants then watched and listened to short video clips of these example sentences twice, focusing their attention on differences between the pronunciation of the target phrases in authentic speech and their expected citation forms (Figure 5.13). In contrast to the rule-based condition, no visual aids (e.g. sound-approximating spellings, colour-coded letters, or punctuation marks) were used to represent changes in pronunciation (cf. Figure 5.5).

Figure 5.13

Sample Slide: Noticing-based Condition – Video Clip Material for Targeted Example

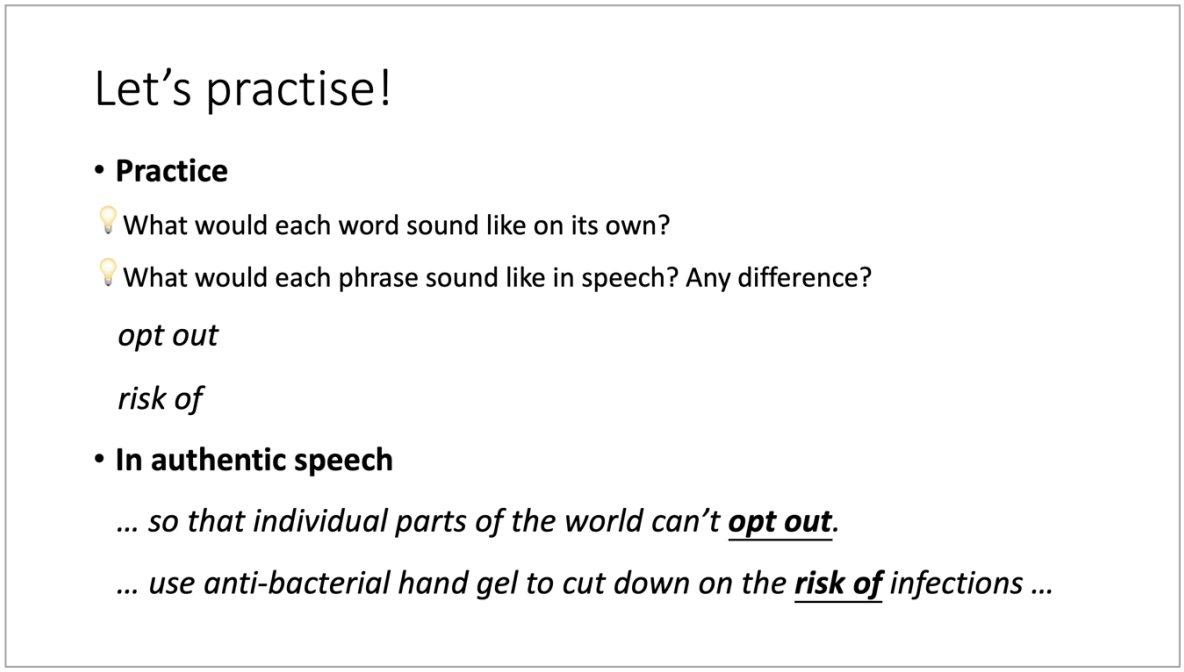


Following these initial examples, participants proceeded to practise with two additional phrases. They were instructed to formulate mental representations of how the words would sound in isolation (i.e. their citation forms) versus how they would sound within the given sentential contexts (Figure 5.14). Participants were allowed to pause the instructional video if additional time was needed for this practice.

Figure 5.14

Sample Slide: Noticing-based Condition – Practice (Compare Isolated vs. Contextualised

Pronunciation)



Participants then watched and listened to short video clips of these phrases within authentic speech contexts (Figure 5.15). Each video clip was played twice while participants were encouraged to replay it as many times as needed. Again, unlike the rule-based condition (cf. Figure 5.8), no visual representations were provided to show the phonological mechanisms involved.

Figure 5.15

Sample Slide: Noticing-based Condition – Practice (Video Clip Material)

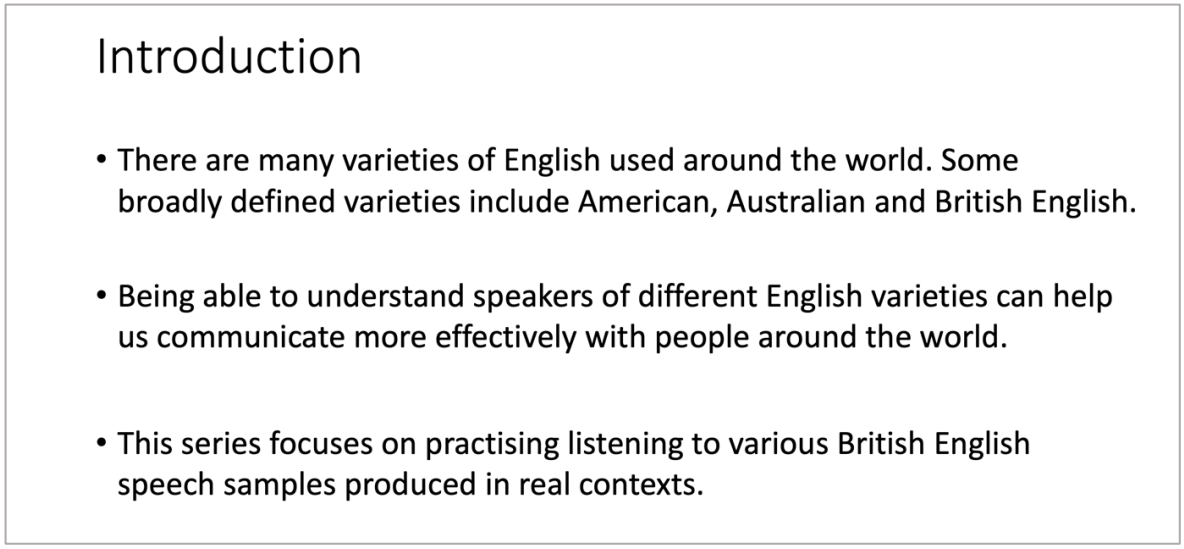
The instruction for each CSP category concluded with the same advanced practice task as used in the rule-based condition, which involved identifying missing words in a sentence. For illustration, refer to Figure 5.9 and Figure 5.10 above.

In summary, the noticing-based condition involved inductive instruction, guiding participants to attend to the differences between the dictionary pronunciation of individual words and their pronunciation in authentic speech contexts, without the use of explicit rule explanations or technical terminology. For each CSP category, participants were presented with five targeted phrases – two initial examples, two practice items, and one item embedded in an advanced word identification task – which were identical to those used in the rule-based condition. This instructional structure was consistently applied across all six CSP categories, although participants in this condition were *not* informed that they were learning distinct 'CSP categories'.

Control Group: Implicit Condition. In contrast to the rule-based and noticing-based conditions, the implicit (control) condition participants did not receive any instruction focused on pronunciation or CSPs. During the introductory session, participants were informed that the intervention programme was designed to enhance their ability to understand speakers of British English, and that they would engage in listening comprehension practice using various speech samples. Figure 5.16 presents a slide from the introductory session outlining the focus of the intervention, illustrating the contrast with the experimental conditions (cf. Figure 5.2 and Figure 5.11).

Figure 5.16

Sample Slide: Implicit (Control) Condition – Introductory Session

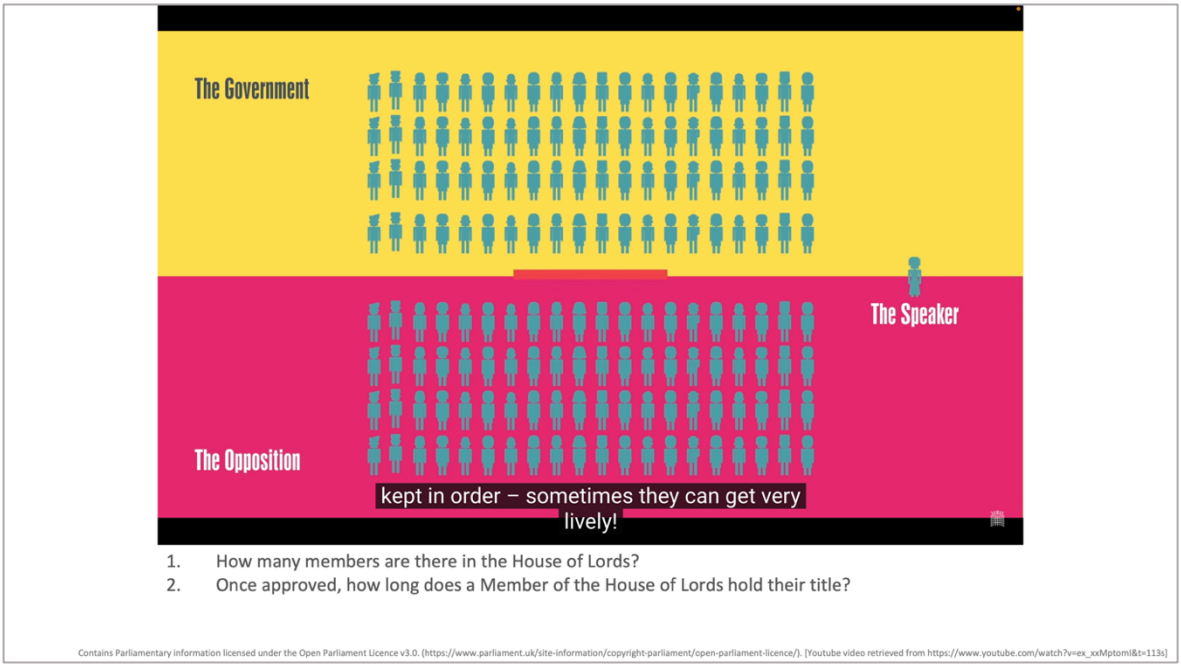


For each CSP category, participants watched and listened to five longer video clips, each approximately one minute in length. These video clips were sourced from the same materials used for instruction in the rule-based and noticing-based conditions, but presented in extended discourse format rather than individual sentences (i.e. incorporating broader contextual information beyond the sentential level). The aim was to provide

participants with meaning-focused listening comprehension practice, instead of directing their attention to specific pronunciation features. Participants were instructed to take notes and answer two pre-designed comprehension questions for each video clip (Figure 5.17).

Figure 5.17

Sample Slide: Implicit (Control) Condition – Extended Video Clip Material

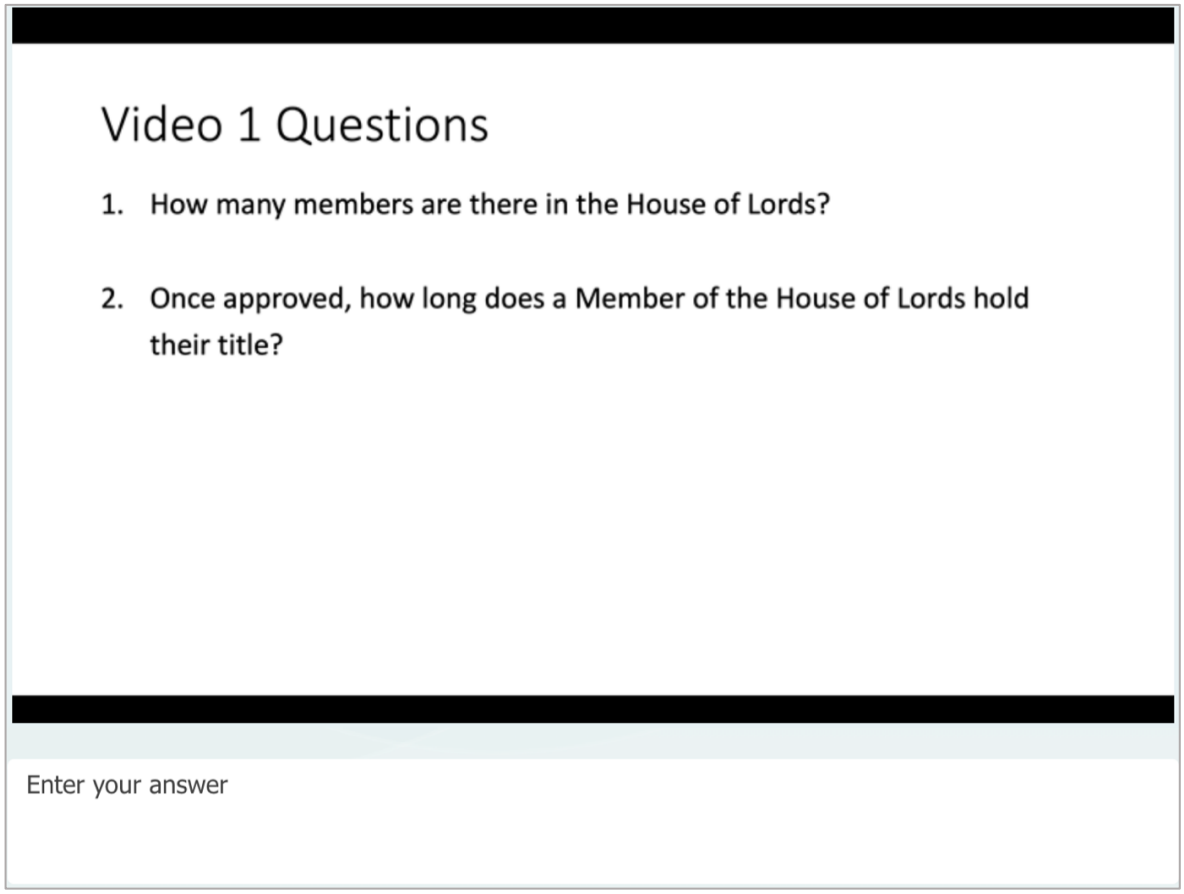


Participants submitted their responses to each comprehension question via corresponding online answer forms, as shown in Figure 5.18. Notably, for each video clip, one of the comprehension questions was strategically designed to require elicitation of information from the sentence containing the target CSP-affected phrase (i.e. the sentence practised by the two experimental groups). For instance, in the example shown, the answer to Question 2 was found in the sentence '*Once approved, … and you are then a Member of the House of Lords, a "peer" for the rest of your life*.' This approach ensured that participants had the opportunity to engage implicitly with the target CSPs (in this case, embedded within the phrase *rest of*), while their focus remained on comprehending the

message meaning rather than on pronunciation (cf. Figure 5.4 for the rule-based instruction and Figure 5.12 for the noticing-based instruction).

Figure 5.18

Sample Slide: Implicit (Control) Condition – Answer Form



Throughout the intervention, participants in the implicit condition were exposed to five examples per CSP category, with the CSP-affected items embedded within extended spoken discourses. However, no pronunciation-related instruction was provided, nor were participants introduced to any CSP terminology or categorisation. They were instructed only to engage in listening comprehension practice. Upon completion of all the intervention components, participants received the answer keys for the comprehension questions.

Summary of Pedagogical Approaches

The three pedagogical approaches described above were implemented to examine whether, and to what extent, participants' ability to process connected speech improved under different intervention conditions, each grounded in distinct theoretical assumptions.

The rule-based condition involved explicit instruction on the rules governing the usage of CSPs, detailing when and how each CSP category occurs. This approach aimed to develop participants' explicit metalinguistic knowledge of CSPs, supporting them in applying these rules to decode novel items in connected speech featuring similar phonological contexts. The noticing-based condition provided induction-oriented, guided instruction, encouraging participants to notice the differences between dictionary pronunciation (i.e. citation form) and real-life pronunciation (i.e. connected speech), without explicit explanations regarding the occurrence of CSPs. It was expected that participants might infer CSP patterns through focused attention to pronunciation changes in the targeted examples, thereby facilitating their processing of novel items in connected speech. The implicit condition, acting as a control group, engaged participants in meaning-focused listening comprehension practice without receiving any pronunciation-related instruction. By strategically incorporating comprehension questions that required attention to sentences containing the CSP-affected phrases, this approach implicitly exposed participants to the target CSPs, while they focused on comprehending the message meaning.

These pedagogical designs provided tightly controlled input and instructional structures, ensuring that any meaningful differences in the pre-test and post-test outcomes could be

more reliably attributed to the effects of the specific intervention conditions. The following section presents the results of this intervention study.

5.2 Results

To evaluate the effectiveness of the interventions, participants' performance on all linguistic measures across the pre-test and post-test was analysed statistically. Prior to examining the principal outcome measures, three supplementary measures – LexTALE, L2 General Listening Comprehension (GLC), and L1 Lexical-Semantic Knowledge Test (LSK) – were assessed using Bayesian regression models. This preliminary analysis aimed to determine whether baseline differences existed between groups at the outset. For clarity, the two experimental groups (i.e. rule-based and noticing-based conditions) are henceforth referred to as the ER (Experimental: Rule-based) and EN (Experimental: Noticing-based) groups, respectively, while the control group (i.e. implicit condition) is referred to as the CT group.

The analysis revealed a group difference between the CT and EN groups in the LexTALE[33], whereas no group differences were found for the GLC and LSK measures. In light of this result, LexTALE scores were included as a covariate in the subsequent models for the principal measures (as will be demonstrated below) to ensure that the intervention effects could be evaluated while controlling for variation in initial English proficiency. Full analyses for the LexTALE and LSK are provided in Appendix F. On the other hand, the GLC results potentially provide another window into the benefits of the

[33] This group difference emerged despite all participants meeting the inclusion criterion of B2-level proficiency or above and being randomly assigned to one of the three conditions (cf. Section 5.1.1). The result was therefore considered incidental.

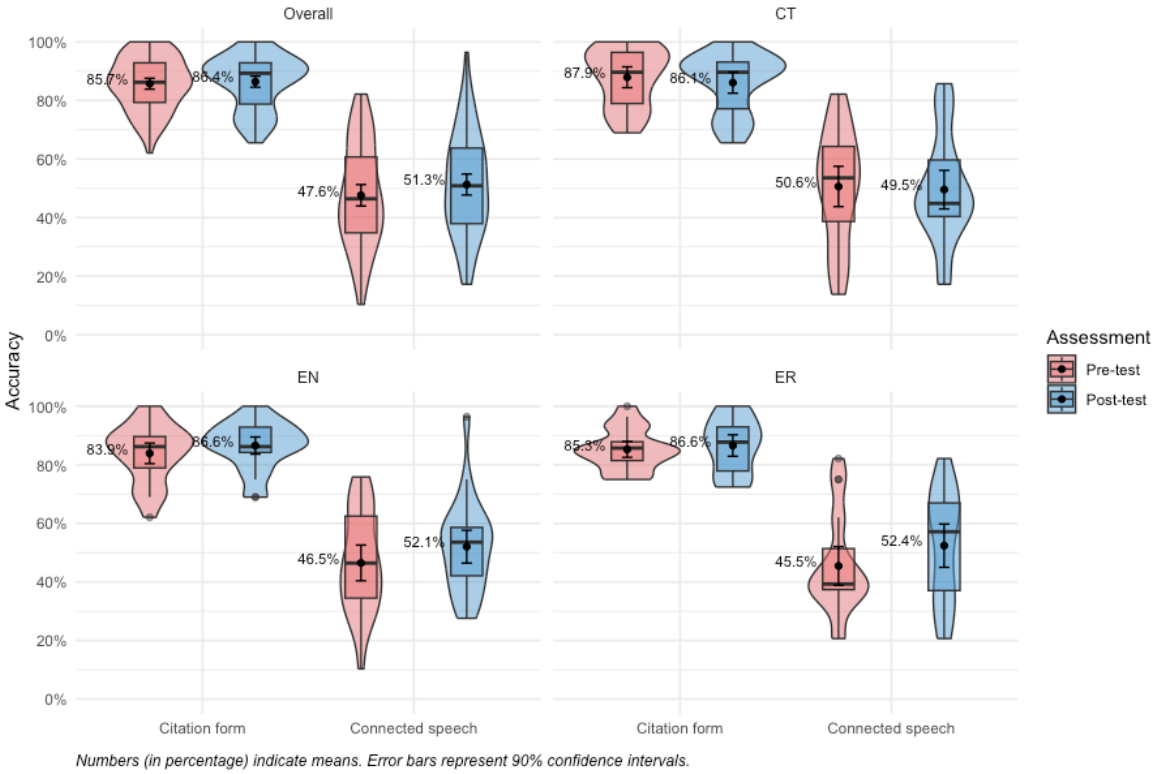
interventions for general listening comprehension, and are therefore reported in the main text following the analyses of principal measures. Lastly, a correlation analysis is presented to explore the interrelationships among all measures.

5.2.1 Connected Speech Dictation vs. Citation Form Dictation

Figure 5.19 presents the accuracy distributions for the connected speech dictation and citation form dictation for each group across the two assessment time points (pre-test and post-test). For comparison, the figure also includes overall performance aggregated across all groups.

Figure 5.19

Connected Speech versus Citation Form Accuracy by Group and Assessment



Overall, participants consistently demonstrated high accuracy in the citation form

condition, achieving 85.7% at the pre-test and 86.4% at the post-test. In contrast, performance in the connected speech condition (where participants processed the same lexical items featuring the target CSPs) was substantially lower, with a mean accuracy of 47.6% at the pre-test and 51.3% at the post-test. These results echoed the outcomes from the Phase 1 study (cf. Section 4.2.1), but with even greater disparities between connected speech and citation form, likely reflecting the slightly lower overall proficiency of participants in this phase. This pattern underscores the persistent challenges faced by L2 participants in processing connected speech compared to citation forms.

In terms of overall improvement from the pre-test to post-test, connected speech showed a numerically larger increase in the mean accuracy (3.7 percentage points) than citation forms (0.7 percentage points). Group-specific performance patterns revealed similar trends, albeit with some variations across groups. The experimental groups (ER and EN), which received pronunciation-focused instruction, exhibited comparatively greater gains in connected speech from the pre-test to post-test. Specifically, the ER group improved from 45.5% to 52.4% (a gain of 6.9 percentage points), and the EN Group from 46.5% to 52.1%. (a gain of 5.6 percentage points). In contrast, the control group (CT), which was exposed only to implicit input through listening comprehension tasks, declined slightly from 50.6% to 49.5% (a decrease of 1.1 percentage points).

Subsequently, these descriptive patterns will be further examined though inferential statistical analyses. Specifically, the forthcoming analyses will focus on the effects of the interventions on connected speech accuracy, addressing two primary questions:

(i) Did participants in each group show improvement in connected speech transcription accuracy following the interventions?

(ii) Were there any group differences indicating greater improvement for one group relative to the others?

The Bayesian Inference Approach

This study employed Bayesian regression models, implemented via the *brms* package in R (Bürkner, 2017), to conduct inferential statistical analyses. The adoption of the Bayesian approach was driven primarily by its improved convergence when fitting complex models that include both fixed and random effects, particularly when working with relatively small sample sizes. As will be demonstrated in the subsequent sections, the models in this study included not only fixed effects for assessments, group contrasts, and their interactions, but also random intercepts and random slopes by participant across assessments. The Bayesian models enabled the inclusion of these effects while achieving successful model convergence for all measures under investigation.

Bayesian regression models share similarities with frequentist mixed effects models in their specification of fixed and random effects, but differ fundamentally in their use of priors – i.e. distributions that reflect prior beliefs about parameter values. These priors serve as initial assumptions, allowing the model to generate plausible parameter estimates based on the observed data. The parameter estimation process is conducted iteratively via Markov Chain Monte Carlo (MCMC) sampling, implemented through the Stan probabilistic programming language (Carpenter et al., 2017). This process evaluates the probability of different parameter values, progressively refining estimates by moving towards higher-probability regions of the parameter space. Model convergence is achieved, typically after thousands of iterations, when the sampled estimates stabilise around specific parameter values. Final estimates are commonly derived from the mean

of these sampling results, known as the posterior distribution. All Bayesian models in this study were run with four chains of 1,500 MCMC iterations (excluding 500 warm-ups per chain), resulting in a total of 6,000 iterations per parameter, unless otherwise specified.

Bayesian model estimates can generally be interpreted in a manner similar to those from frequentist models. In fact, when non-informative priors (e.g. flat priors) are specified – i.e. allowing the model to explore a wide range of values based primarily on the observed data without providing strong a priori hypotheses – the estimates obtained from the Bayesian approach are highly comparable to those from the frequentist approach. However, Bayesian models differ particularly in their treatment of uncertainty. Unlike frequentist models, which rely on *p*-values, Bayesian models provide credible intervals (CIs) derived from the posterior distribution. These CIs represent the range within which true parameter values are likely to fall. This contrasts with the frequentist approach, where *p*-values do not directly reflect the probability of specific parameter values. For comparative purposes, while this study adopted the Bayesian approach, results from frequentist mixed effects models are also reported in Appendix G, where applicable.

Bayesian Logistic Mixed Effects Regression Model

To analyse the effects of Group (with three levels: ER, EN, CT), Assessment (with two levels: pre-test, post-test), and their interactions on connected speech transcription accuracy, a Bayesian logistic mixed effects regression model was constructed. The model was fitted with a random effect structure of (Assessment.ct | PID), specifying a random intercept for each participant (PID) and a random slope for each participant's performance across assessments (centre coded as Assessment.ct, with pre-test = -0.5, post-test = 0.5).

Additionally, given the observed baseline differences in the LexTALE at the pre-test (cf. Appendix F), LexTALE scores were included as a covariate to control for variation in initial proficiency. The scores were centred around the mean (hence labelled as LexTALE.ct), thereby allowing the model intercept to represent the grand mean (i.e. average proficiency) rather than the value associated with a raw score of zero.

Group contrasts were coded using the simple coding scheme, and two versions of the model were run: one specifying CT as the reference level and the other specifying ER as the reference level. This coding generated two contrasts: CT_VERSUS_ER and CT_VERSUS_EN in Version 1, and ER_VERSUS_CT and ER_VERSUS_EN in Version 2, thereby capturing all pairwise group contrasts (Table 5.5). The sum-to-zero contrast coding ensured that the model intercept represented an estimate of the grand mean rather than the mean of a specific reference level.

Table 5.5

Simple Coding Scheme for Intervention Groups in Regression Models

	Contrast I: CT_VERSUS_ER	**Contrast II: CT_VERSUS_EN**
Control (CT) = reference	-1/3	-1/3
Rule-based (ER)	2/3	-1/3
Noticing-based (EN)	-1/3	2/3
	Contrast I: ER_VERSUS_CT	**Contrast II: ER_VERSUS_EN**
Control (CT)	2/3	-1/3

Rule-based (ER) = reference	-1/3	-1/3
Noticing-based (EN)	-1/3	2/3

Model for Connected Speech: Version 1

$$Connected\ speech \sim CT\_VERSUS\_ER * Assessment.ct + CT\_VERSUS\_EN * Assessment.ct + LexTALE.ct + (Assessment.ct \mid PID)$$

Model for Connected Speech: Version 2

$$Connected\ speech \sim ER\_VERSUS\_CT * Assessment.ct + ER\_VERSUS\_EN * Assessment.ct + LexTALE.ct + (Assessment.ct \mid PID)$$

The two versions of the model complementarily encompassed all three group contrasts (i.e. CT_VERSUS_ER, CT_VERSUS_EN, ER_VERSUS_EN) and their respective interactions with Assessment.ct. The fixed-effect coefficients from the two versions of the model are summarised in Table 5.6. In addition, a similar model without contrast coding was run to explore the effects of Assessment.ct within each group individually, reported as Assessment.ct (CT/ER/EN) under the overall main effect (Assessment.ct) in Table 5.6. The column 'B<>0' represents the posterior probability that each parameter is either above or below zero, depending on the direction of the estimate. For instance, for a positive estimate, B<>0 reflects the probability that the parameter is positive, and vice versa for a negative estimate. Practically, when flat priors are applied (as in this model), the probability that the parameter differs from zero (i.e. 1 minus B <> 0) can be interpreted as analogous to a one-tailed *p*-value in the frequentist approach (Marsman & Wagenmakers, 2017). These probabilities are visualised in Figure 5.20 through the posterior density distributions for the main effect and key interaction parameters (cf. one-tailed *p*-values from frequentist mixed effects models in Appendix G).

Credible intervals (CIs) were used to evaluate evidence against a null hypothesis (equivalent to *p*-value = .05 in frequentist terms). Specifically, if the CIs did not cross zero, the null hypothesis of no difference would be rejected. Note that 95% CIs were used as the default for the main model terms, while 90% CIs were applied to parameters with one-tailed hypotheses (in boldface): the main effects (Assessment.ct) and interactions involving contrasts between the control and experimental groups (CT_VERSUS_ER:Assessment.ct, CT_VERSUS_EN:Assessment.ct). This reflects our directional hypotheses of (1) positive main effects Assessment.ct, indicative of stronger rather than weaker performance following the pedagogical interventions; and (2) greater improvement for the two experimental groups (ER and EN), which received pronunciation-focused instruction, compared to the control group (CT), which was exposed only implicitly to target examples amongst a larger sample of input.

Table 5.6

Summary of Bayesian Model Output for Connected Speech Dictation

	Estimate	**Est. Error**	**CI_lower**	**CI_upper**	**B<>0**
Intercept	-0.03	0.07	-0.15	0.10	0.65
LexTALE.ct	3.54	0.56	2.47	4.64	1
Assessment.ct	**0.17**	**0.10**	**0.01**	**0.33**	**0.96**
Assessment.ct (CT)	**-0.03**	**0.18**	**-0.32**	**0.26**	**0.58**
Assessment.ct (ER)	**0.29**	**0.21**	**-0.05**	**0.64**	**0.92**
Assessment.ct (EN)	**0.24**	**0.13**	**0.02**	**0.46**	**0.97**
CT_VERSUS_ER	0.04	0.17	-0.28	0.36	0.60

CT_VERSUS_EN	0.24	0.17	-0.09	0.57	0.92
ER_VERSUS_EN	0.19	0.16	-0.13	0.51	0.88
CT_VERSUS_ER:Assessment.ct	**0.33**	**0.24**	**-0.07**	**0.73**	**0.91**
CT_VERSUS_EN:Assessment.ct	**0.30**	**0.23**	**-0.09**	**0.66**	**0.90**
ER_VERSUS_EN:Assessment.ct	-0.03	0.24	-0.51	0.43	0.55

Note. Estimates are in log-odds space. For rows in bold, CIs are one-tailed (as we have directional predictions), otherwise they are two-tailed.

The model intercept (b = -0.03) reflects the fact that overall connected speech accuracy, averaged across groups and assessments, is slightly below 50% (0 in log-odds), yet statistical uncertainty is manifested in the CIs[-0.15, 0.10] crossing zero. The estimate for LexTALE.ct (b = 3.54) indicates a strong positive association between higher proficiency and higher connected speech accuracy, with the CIs[2.47, 4.64] showing substantial certainty.

The group contrasts (i.e. CT_VERSUS_ER, CT_VERSUS_EN, ER_VERSUS_EN) represent the difference between groups averaged across the pre-test and post-test. However, these averages per se do *not* reveal the intervention effects, as they do not account for changes over time. To evaluate the intervention effects, the critical parameters of interest are: (a) the main effects of Assessment.ct, capturing overall and group-specific improvements from the pre-test to post-test; and (b) the interactions between group contrasts and Assessment.ct, capturing differential improvements across intervention groups.

The main effect of Assessment.ct (b = 0.17, CIs[0.01, 0.33]) indicates an overall increase

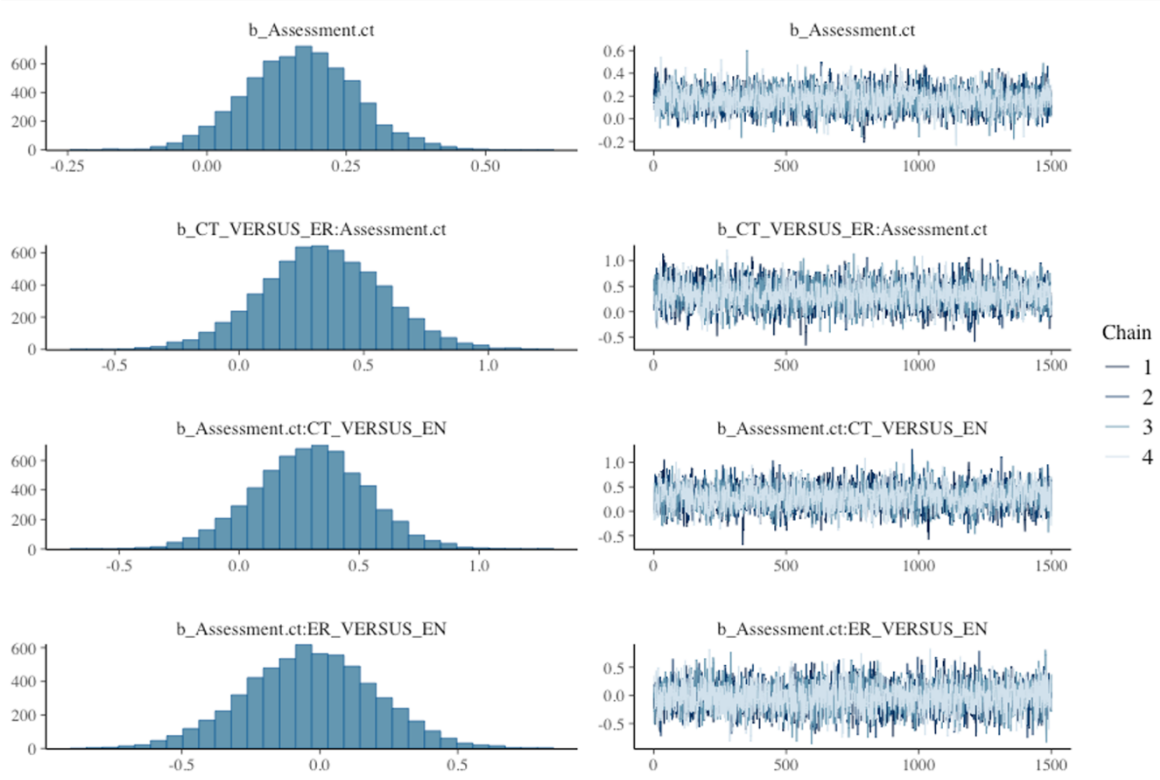
in connected speech accuracy (averaged across groups) from the pre-test to post-test. Group-specific effects show differing trends. A near-zero estimate for the CT group (b = -0.03, CIs[-0.32, 0.26]) with CIs crossing zero indicates no evidence of a meaningful difference from the pre-test to post-test (in fact, a very minor decrease, as shown by the negative estimate and the visual pattern in Figure 5.19). Contrastively, positive estimates for the EN group (b = 0.24, CIs[0.02, 0.46]) and the ER group (b = 0.29, CIs[-0.05, 0.64]) suggest improvement, though the CIs for the latter indicate some degree of uncertainty as they slightly cross zero. Nevertheless, while there are positive trends towards greater improvement for the experimental groups compared to the control group, the interactions between group contrasts and Assessment.ct show statistical uncertainty, with the CIs crossing zero for CT_VERSUS_ER:Assessment.ct (b = 0.33, CIs[-0.07, 0.73]) and CT_VERSUS_EN:Assessment.ct (b = 0.30, CIs[-0.09, 0.66]). ER_VERSUS_EN:Assessment.ct (b = -0.03, CIs[-0.51, 0.43]) similarly indicates no evidence of differential improvement between the ER and EN groups, with the CIs broadly distributed around zero.

Figure 5.20 presents the Bayesian posterior density distributions for the overall main effect and interaction terms (left), along with their corresponding MCMC sampling trace plots (right), which document the progression and convergence of the model. In the density distribution plots, the x-axis represents parameter estimates, the y-axis represents the density of the iterations. The proportion of the posterior distribution lying above or below zero (depending on the direction of the estimated effect) corresponds to the B<>0 values in Table 5.6. For example, 96% of the iterations for Assessment.ct yield positive estimates. Similarly, 91% of the iterations for CT_VERSUS_ER:Assessment.ct and 90% for CT_VERSUS_EN:Assessment.ct are above zero. In contrast, 55% of the iterations for

ER_VERSUS_EN:Assessment.ct fall below zero, consistent with its negative model estimate. It is important to note that while these proportions – with zero set as the reference point – are particularly informative for evaluating the likelihood of the direction of effect (i.e. positive versus negative), they do not indicate the plausibility of specific parameter values.

Figure 5.20

Bayesian Posterior Density Distributions for Key Parameters in Connected Speech Model



Note. The trace plots on the right illustrate four MCMC chains of 1,500 iterations (totalling 6,000 iterations) for each parameter. The substantial overlap (white regions) and consistent oscillation around a stable mean indicate that the sampler has stabilised and the model has achieved convergence.

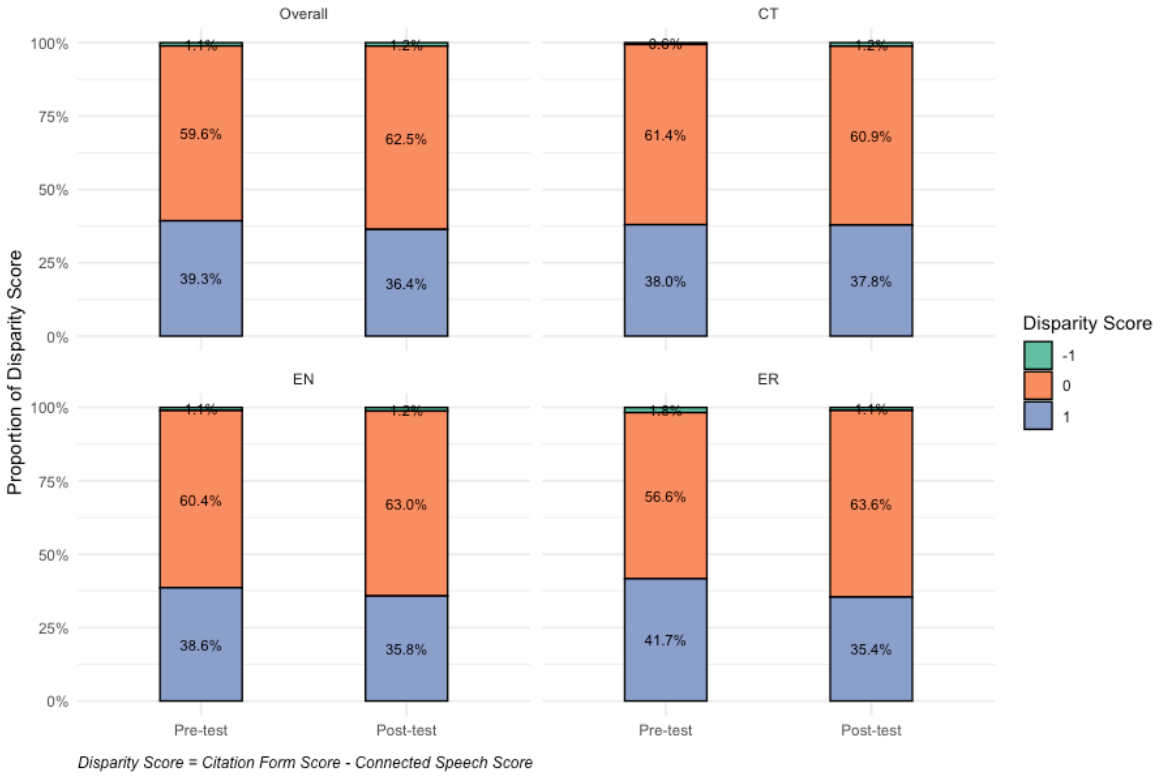
Critically, while the analysis of connected speech accuracy is straightforward and widely employed in prior studies, it may not fully capture CSP-related difficulties. Specifically, lower performance on connected speech items alone does not necessarily imply difficulties with CSPs if participants also struggle with the citation form of the same items. A more precise measure involves assessing the disparity between participants' responses to connected speech and citation form. Therefore, we repeat the analyses using the disparity score metric, and the results are presented in the subsequent section.

5.2.2 Disparity Score

A disparity score was calculated by subtracting the connected speech score from the citation form score for each pair of responses (i.e. one participant's scores for the same stimulus item in connected speech and citation form). Given the binary nature of the outcomes in each measure (1 = accurate response, 0 = inaccurate response), the disparity score could take one of three possible values: -1, 0, or 1. A score of 1 indicates accurate transcription of the citation form but not the same item in connected speech; a score of 0 indicates no difference; a score of -1 indicates accurate transcription in connected speech but not in the citation form. Figure 5.21 presents the proportions of disparity scores across groups and assessment time points.

Figure 5.21

Proportions of Disparity Scores by Group and Assessment



It can be observed that a disparity score of -1 (i.e. accurate transcription in connected speech but inaccurate transcription in the corresponding citation form) is rare, representing approximately 1% of the data. Such low proportions could be considered noise (given that the citation form is generally easier to process for L2 learners), likely reflecting factors unrelated to CSPs, such as participants' familiarity with certain connected speech items due to prior or frequent exposure in their language experience. These cases were therefore excluded from further analysis.

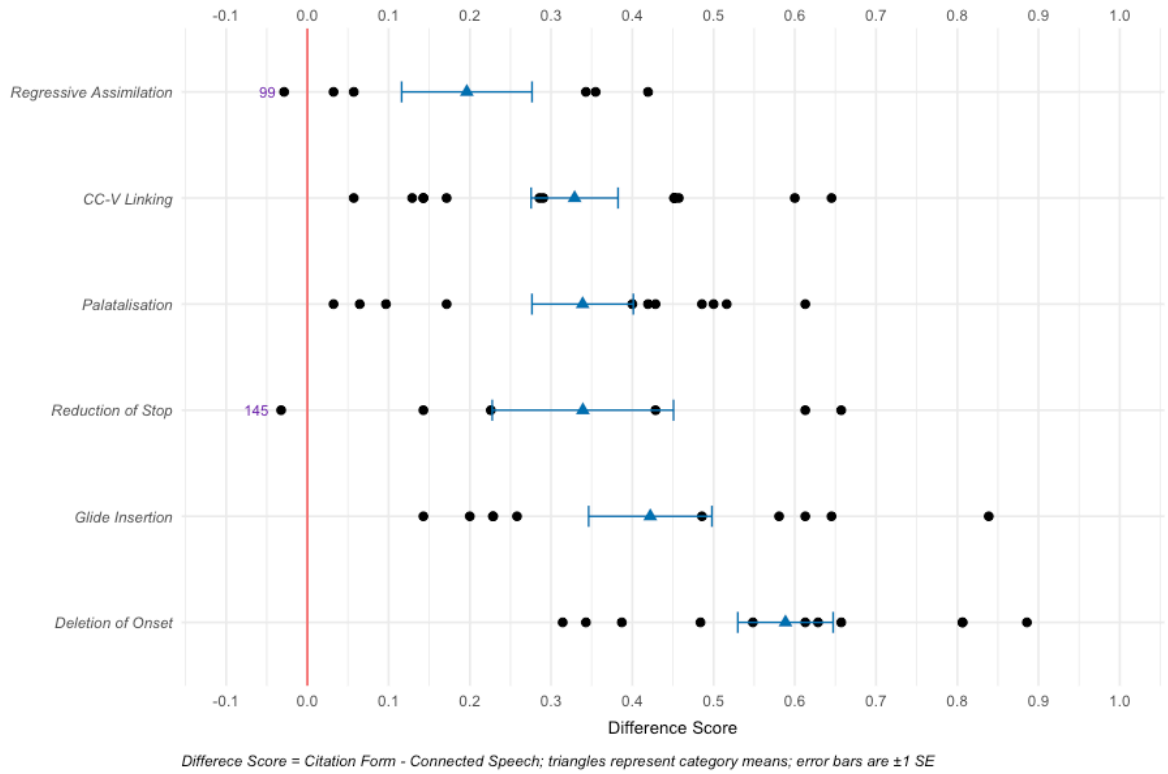
This exclusion enabled the application of logistic regression models with binary outcomes (0 and 1), aligning with the analysis of connected speech accuracy (cf. Section 5.2.1) and allowing for direct comparison of results from two perspectives: (1) increased accuracy in transcribing connected speech items; and (2) decreased CSP-induced difficulty as

indicated by the disparity score. Note that with cases of -1 excluded, subsequent analyses essentially estimate the likelihood of CSPs causing difficulty versus the likelihood of CSPs having no impact. To facilitate interpretation, the binary outcomes were recoded such that a value of 1 indicates *no* disparity (i.e. no CSP-specific difficulty) and 0 indicates disparity (i.e. CSP-specific difficulty). This recoding ensures that higher scores reflect stronger performance, with positive slopes representing improvement over time, consistent with the interpretation of connected speech accuracy reported above.

Additionally, a 'difference score' was computed for each stimulus item at the pre-test by subtracting the connected speech accuracy from the citation form accuracy. Figure 5.22 presents these difference scores by CSP category, along with category means (shown as triangles) and the corresponding standard errors. Following the same procedure as in the Phase 1 study (cf. Section 4.2.2 and Figure 4.7), items positioned to the left of the red line – i.e. those with lower accuracy in citation form than in connected speech (n = 2) – were excluded. These two items were considered unsuitable for evaluating CSP-specific difficulty, as they were already easier to process in connected speech prior to the interventions, and the focus of this study is to explore the benefits of the interventions for items negatively affected by the presence of CSPs.

Figure 5.22

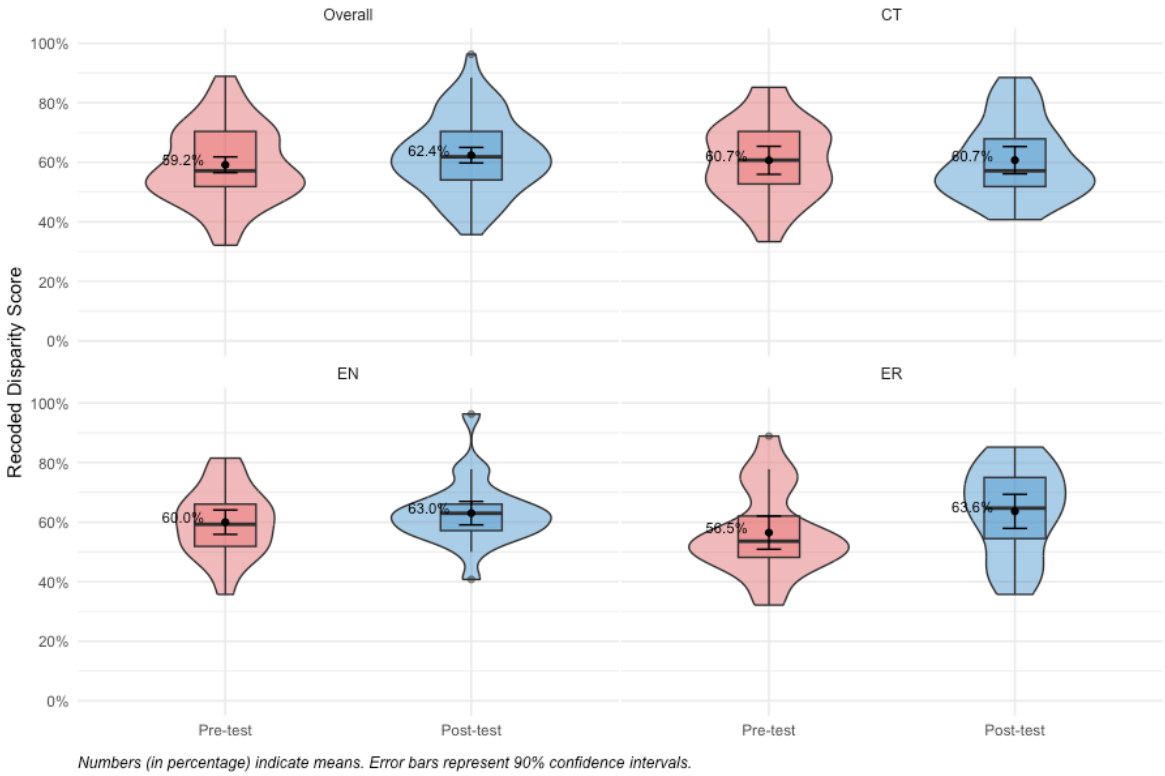
Mean Difference Scores for Individual Stimulus Items at Pre-test



The two rounds of exclusion finalised the data cleaning for the disparity score metric. Figure 5.23 presents the final, recoded binary disparity scores by group and assessment. The percentages indicate the proportion of items with *no* difficulty (i.e. score = 1 in the recoded variable), such that higher percentages reflect decreased CSP-specific difficulty, and therefore, stronger performance.

Figure 5.23

Recoded Disparity Scores by Group and Assessment



Note. Higher percentages reflect stronger performance, with 100% indicating no CSP-specific difficulty in processing connected speech.

Bayesian Logistic Mixed Effects Regression Model

A Bayesian logistic mixed effects regression model was fitted with disparity scores as the dependent variable, adopting the same fixed and random effect structures used in the connected speech accuracy model. As before, two versions of the model were run to encompass all three pairwise group contrasts:

Model for Disparity Scores: Version 1

$$Disparity\ score \sim CT\_VERSUS\_ER * Assessment.ct + CT\_VERSUS\_EN * Assessment.ct + LexTALE.ct + (Assessment.ct \mid PID)$$

Model for Disparity Scores: Version 2

$$Disparity\ score \sim ER\_VERSUS\_CT * Assessment.ct + ER\_VERSUS\_EN * Assessment.ct + LexTALE.ct + (Assessment.ct \mid PID)$$

Table 5.7 summarises the estimates and standard errors for the target parameters: the intercept, LexTALE.ct, overall and group-specific effects Assessment.ct, and interaction terms. As in the previous analysis, 95% CIs are reported for two-tailed parameters, while 90% are used for parameters with one-tailed directional hypotheses (in boldface).

Table 5.7

Summary of Bayesian Model Output for Disparity Scores

	Estimate	**Est. Error**	**CI_lower**	**CI_upper**	**B<>0**
Intercept	0.45	0.04	0.36	0.54	1
LexTALE.ct	2.22	0.38	1.48	2.96	1
Assessment.ct	**0.15**	**0.09**	**0.01**	**0.29**	**0.96**
Assessment.ct (CT)	**0.01**	**0.16**	**-0.25**	**0.27**	**0.52**
Assessment.ct (ER)	**0.32**	**0.17**	**0.04**	**0.60**	**0.97**
Assessment.ct (EN)	**0.13**	**0.13**	**-0.08**	**0.35**	**0.86**
CT_VERSUS_ER:Assessment.ct	**0.31**	**0.20**	**-0.02**	**0.65**	**0.94**
CT_VERSUS_EN:Assessment.ct	**0.13**	**0.20**	**-0.20**	**0.46**	**0.75**
ER_VERSUS_EN:Assessment.ct	-0.17	0.21	-0.58	0.23	0.81

Note. Estimates are in log-odds space.

Note that the intercept and the corresponding CIs are positive (b = 0.45, CIs[0.36, 0.54]), reflecting greater odds of experiencing no difficulty than experiencing difficulty, consistent with the patterns depicted in Figure 5.21. The positive estimate for LexTALE.ct

(b = 2.22, CIs[1.48, 2.96]) with the CIs not crossing 0, indicates that higher proficiency is strongly associated with reduced difficulty with CSPs.

The main effect of Assessment.ct (b = 0.15, CIs[0.01, 0.29]) indicates an overall improvement (averaged across groups) from the pre-test to post-test. Differing trends can be observed among individual groups. Similar to connected speech accuracy, the CT group shows a near-zero estimate (b = 0.01, CIs[-0.25, 0.27]), indicating no evidence of change over time and reflecting the identical pre-test and post-test means shown in Figure 5.23. On the other hand, both the ER group (b = 0.32, CIs[0.04, 0.60]) and the EN group (b = 0.13, CIs[-0.08, 0.35]) demonstrate positive estimates, though the CIs for the latter cross zero, indicating greater uncertainty. Notably, this pattern contrasts with the connected speech accuracy analysis, where the EN group shows greater certainty with the CIs not crossing zero (cf. Table 5.6).

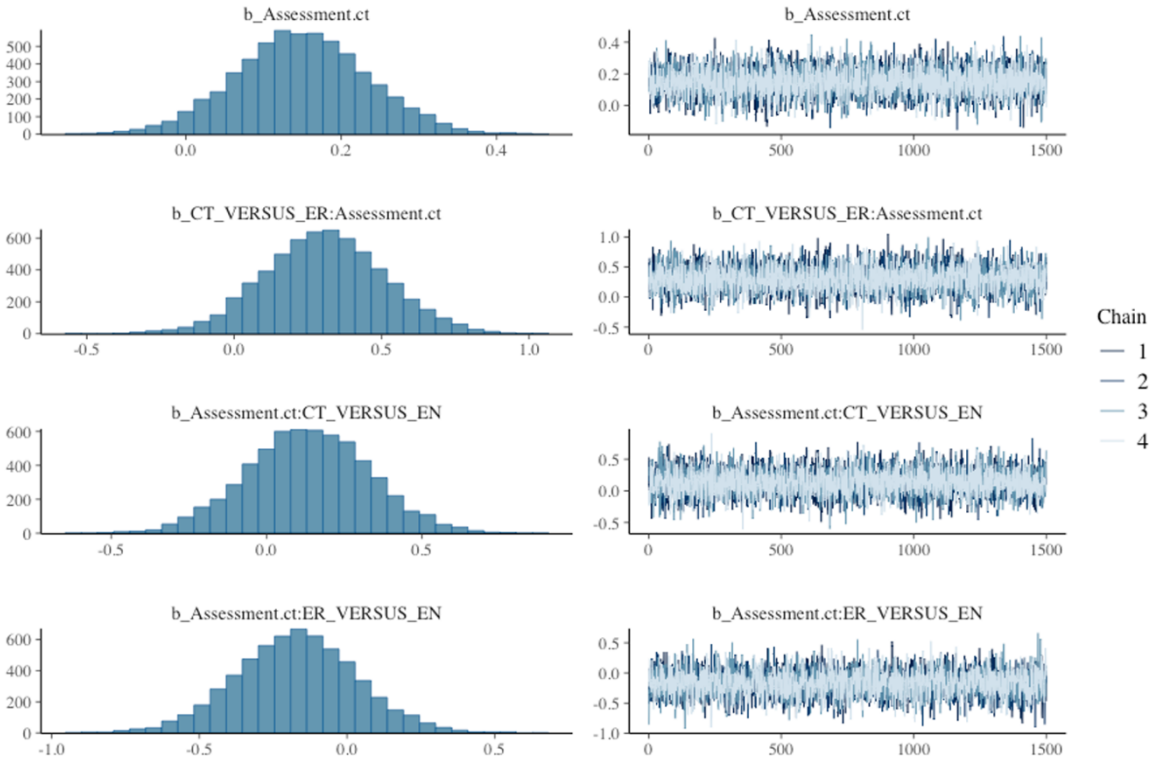
However, despite the positive trends exhibited by the experimental groups relative to the control group, the corresponding interaction terms again yield CIs that cross zero: CT_VERSUS_ER:Assessment.ct (b = 0.31, CIs[-0.02, 0.65]) and CT_VERSUS_EN:Assessment.ct (b = 0.13, CIs[-0.20, 0.46]). Similarly, the contrast between the two experimental groups, ER_VERSUS_EN:Assessment.ct (b = -0.17, CIs[-0.58, 0.23]), provides no evidence of differential improvement. Thus, while numerical trends indicate greater improvement for the experimental groups, the interaction terms provide no robust statistical evidence of group differences.

Figure 5.24 presents the posterior density distributions for the main effect and interaction parameters, along with their corresponding MCMC sampling trace plots. Corresponding

to the B<>0 values reported in Table 5.7, 96% of the iterations for Assessment.ct yield positive estimates, 94% for CT_VERSUS_ER:Assessment.ct, and 75% for CT_VERSUS_EN:Assessment.ct. For ER_VERSUS_EN:Assessment.ct, 81% of the iterations are negative due to its negative estimate. As mentioned earlier, while these distributions indicate the likelihood of the direction of effects, they do not endorse any specific parameter values.

Figure 5.24

Bayesian Posterior Density Distributions for Key Parameters in Disparity Score Model



While these results provide valuable insights into the (un)certainty associated with each parameter, they do not directly suggest whether the data is more likely under H_1 or under H_0. Critically, although CIs determine whether the null could be rejected (i.e. whether there is support for an effect), they do not evaluate the plausibility of H_0 per se. In other

words, where CIs include zero – as with the interaction terms in this study – they do not tell us whether there is evidence for H_0 over H_1 or whether the evidence is ambiguous and thus cannot distinguish between the competing hypotheses. This limitation is also true of frequentist *p*-values exceeding .05.

Therefore, the subsequent section employs a different inferential statistic – the Bayes factor – to quantify the relative evidence for H_0 versus H_1 for the key interaction terms. Unlike conventional statistical thresholds, Bayes factors allow us to differentiate between (i) evidence in favour of H_1, (ii) evidence in favour of H_0, and (iii) ambiguous evidence. This approach serves to offer a more nuanced evaluation of potential group differences.

Bayes Factor Analysis

For the purposes mentioned above, Bayes factors were computed for the key interaction terms:

(a) CT_VERSUS_ER:Assessment.ct

(b) CT_VERSUS_EN:Assessment.ct

(c) ER_VERSUS_EN:Assessment.ct

A Bayes factor (BF) quantifies the relative likelihood of the observed data under two competing hypotheses: H_1 (representing the presence of an effect) and H_0 (representing the null or zero effect). In this study, BFs were computed following Dienes's (2008) approach (the R code is included in Appendix H). This calculator requires, for each hypothesis to be tested: (1) two numbers which represent the data – a sample estimate and its standard error; and (2) a predicted effect (denoted as *x*) which reflects a prior belief about the target parameter. The calculator then evaluates the likelihood of the data under

H_0 (represented as a point null) versus under H_1 (modelled as full or half normal for one or two-tailed tests, respectively) with a mean of 0 and a standard deviation set to the predicted effect *x*. This distribution for H_1 captures the fact that smaller effects are generally more probable than larger ones, thereby imposing a more stringent structure than assigning *x* as the mean of H_1, wherein some larger effects would be equally probable as smaller ones symmetrically around *x*. A Bayes factor can be expressed mathematically as:

$$BF_{10} = \frac{P(Data|H_1)}{P(Data|H_0)} = \frac{Likelihood\ under\ H_1}{Likelihood\ under\ H_0}$$

Interpretatively, a BF of 1 indicates that the data is equally likely under H_1 and H_0. A BF larger than 1 indicates evidence in favour of H_1 – for example, a BF of 3 means the data is three times more likely under H_1 than under H_0. Conversely, a BF smaller than 1 indicates evidence in favour of H_0 – for example, a BF of 0.33 (one-third) means the data is three times more likely under H_0 than under H_1.

For each parameter, the sample estimate and standard error are retrieved from the relevant coefficients in the Bayesian logistic mixed effects regression model[34] (as reported in Table 5.7). These values are expressed in log-odds space, meeting the assumptions of

[34]An alternative approach, which can be implemented within *brms*, is to compute a Bayes factor based on comparing two full mixed effects models (Schad et al., 2023): (1) a model including the parameter of interest (similar to the model fitted above, but critically with an informed prior specifically for the parameter of interest); and (2) a model where that parameter is excluded, meaning it is set to zero. The probability of the data under each model is then approximated using bridge sampling (Gronau et al., 2020). Silvey et al. (2024) conducted a partial check on whether these conceptually similar approaches produced similar results when computed over the same datasets (and using the same informed priors as the model of H_1) and found results that were qualitatively highly similar. Here we use the Dienes's (2008) approach on the grounds that it is computationally easier and faster.

normality assumed by the calculator. The determination of an appropriate predicted effect requires cautious consideration. Ideally, we would use values from a similar previous study or pilot. However, in the absence of this, in this study, a motivated-maximum approach is employed following Silvey et al. (2024). This approach involves estimating the maximum effect that may be reasonably expected and using half of this value as the predicted effect. The rationale is that that under a normal distribution, the maximum is approximately two standard deviations above the mean, and thus if the predicted effect is one standard deviation, it is possible to work backwards and obtain this value by halving the maximum.

Specifically, to calculate predicted effects (x) for the interaction between each of the three group contrasts and assessment, we used an estimate of the average effect of assessment (in log-odds) for the two groups in question. For example, for CT_VER-SUS_ER:Assessment.ct, we used an estimate of the average effect of assessment across the CT and ER groups. Since in the original full model (Table 5.7) the main effect (Assessment.ct) represents the overall improvement across all three groups, we fitted three additional models to isolate the corresponding average effects – each including a contrast of two groups of interest (i.e. CT/ER, CT/EN, ER/EN) – using a centred coded group variable (Group.ct) while retaining the same random effect structure as in the full model. The resulting estimates for the average effects of assessment were:

(a) CT and ER groups: 0.16

(b) CT and EN groups: 0.07

(c) ER and EN groups: 0.22

This method for estimating the effect size for an interaction in a two-by-two design is

based on the notion that the maximum interaction would be seen if one group showed the entirety of the effect in question (here the effect of learning) and the other group showed no effect (zero improvement), and a predicted effect can be obtained by working backwards from that. The full logic is laid out in Appendix I.

Table 5.8 summarises the sample estimates, standard errors, predicted effects, and the resulting BFs for the three interaction parameters. Consistent with model estimation, one-tailed tests were specified for CT_VERSUS_ER:Assessment.ct and CT_VERSUS_EN:Assessment.ct (in boldface), where directional hypotheses were made (i.e. positive effects were expected). A two-tailed test was used for ER_VERSUS_EN:Assessment.ct, for which no directional hypothesis was posited. Unlike *p*-values, BFs provide a continuous scale of evidential strength (i.e. the higher a BF, the stronger evidence for H_1, and vice versa for H_0). However, BFs can also be interpreted using threshold criteria. We adopted those proposed by Jeffreys' (1961) (see Table 5.9). Additionally, we report robustness regions, which specify the range of predicted thresholds that would yield the same evidential category (i.e. between 1/3 and 3 in each case).

Table 5.8

Summary of Bayes Factors for Interaction Parameters for Disparity Scores

	Sample Estimate	**Sample SE**	**Predicted Effect**	**Bayes Factor**	**Robustness Region**
CT_VERSUS_ER:Assessment.ct	**0.31**	**0.20**	**0.16**	**2.08**	**[0.01, 3.77]**
CT_VERSUS_EN:Assessment.ct	**0.13**	**0.20**	**0.07**	**1.13**	**[0.01, 1.07]**
ER_VERSUS_EN:Assessment.ct	-0.17	0.21	0.22	0.82	[-0.84, 0.84]

Note. The first three columns represent inputs to the calculator. The final column provides the range of values of predicted effects which would produce the same qualitative result (i.e. a BF between 1/3 and 3 indicating ambiguous evidence).

A $BF_{(0,0.16)}$ of 2.08 for CT_VERSUS_ER:Assessment.ct suggests that the data is approximately twice as likely under H_1 (i.e. the ER group improved more than the CT group) than under H_0 (i.e. no differential improvement). While this BF leans towards the positive direction in favour of H_1, the evidence falls within the ambiguous region, meaning it cannot be fully differentiated from H_0. For CT_VERSUS_EN:Assessment.ct, a $BF_{(0,\,0.07)}$ of 1.13 suggests that the data is almost equally likely under H_1 and H_0, meaning that the evidence provides equivalent support for the presence of an interaction (H_1) and the absence of an interaction (H_0). Finally, a $BF_{(0,0.22)}$ of 0.82 for ER_VERSUS_EN:Assessment.ct slightly leans towards H_0 but again remains within the ambiguous region.

Table 5.9

Interpretation Scale for Bayes Factor (Jeffreys, 1961)

BF_{10}			Interpretation
	>	100	Extreme evidence for H_1
30	–	100	Very strong evidence for H_1
10	–	30	Strong evidence for H_1
3	–	10	Moderate evidence for H_1
1	–	3	Anecdotal evidence for H_1
	1		No evidence for H_1 or H_0
1/3	–	1	Anecdotal evidence for H_0

1/10	–	1/3	Moderate evidence for H_0
1/30	–	1/10	Strong evidence for H_0
1/100	–	1/30	Very strong evidence for H_0
	<	1/100	Extreme evidence for H_0

Note. The threshold of BF > 3 for moderate/substantial evidence generally corresponds to $p < .05$ in frequentist terms, though this equivalence is not guaranteed.

Overall, the Bayes factor results suggest that the lack of group-by-assessment interaction effects in the Bayesian model (Table 5.7) cannot be interpreted as support for the null hypothesis (i.e. that there is no differential improvement). Instead, all three BFs fall within the ambiguous region, indicating that the data does not provide clear evidence either for or against the presence of interaction effects. Although the evidence trends towards the predicted direction for the contrast between the CT and ER groups, the strength of this evidence remains inconclusive.

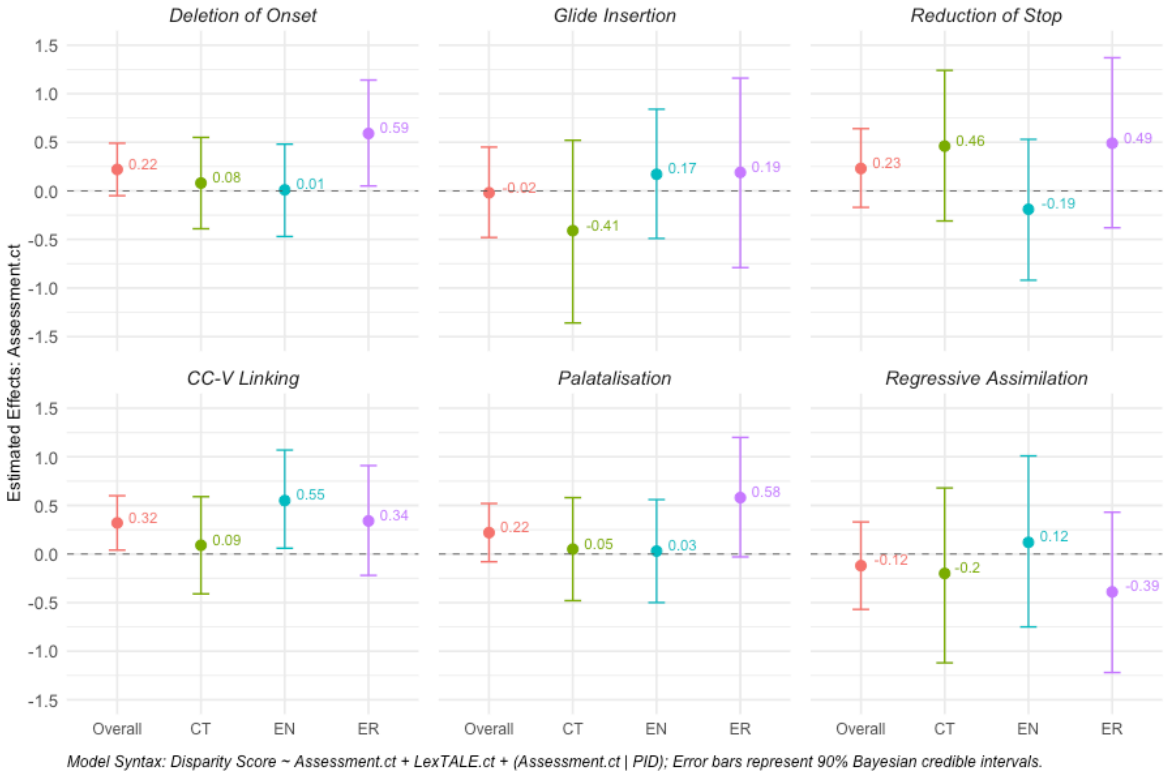
5.2.3 Disparity Score by CSP Category

In light of the prior analyses, where the inferential statistics yielded largely ambiguous evidence even when looking across the whole dataset (cf. Table 5.7 and Table 5.8), we decided not to conduct further inferential statistics for the subsets of individual CSP categories. However, we pursued exploratory modelling for visualisation purposes. Specifically, for each CSP category, three Bayesian logistic mixed effects models were fitted to estimate the effect of Assessment.ct within each group (CT, EN, ER). The resulting estimates and their corresponding 90% CIs are displayed in Figure 5.25, while full model outputs are included in Appendix J.

To ensure convergence in these smaller subset models while maintaining a consistent random effect structure, the number of MCMC sampling iterations was increased to 3,500 per chain (excluding 500 warm-ups). This operation exemplifies a practical advantage of Bayesian modelling – namely, the ability to facilitate model convergence by increasing the number of sampling iterations as needed, as discussed earlier (cf. Section 5.2.1: *The Bayesian Inference Approach*).

Figure 5.25

Estimated Intervention Effects for Disparity Scores by Group and CSP Category



Note. The six CSP categories are arranged in descending order of difficulty based on the Phase 1 results (cf. Figure 4.8) – beginning with the most challenging category, *Deletion of Onset* (upper left), followed by *Glide Insertion*, *Reduction of Stop*, *CC-V Linking*, *Palatalisation*, and ending with the least challenging category, *Regressive Assimilation* (lower right). Estimates represent the effect of assessment (Assessment.ct), indicating

changes from the pre-test to post-test for each group. The overall effect (aggregated across all groups) is also included in each panel for reference.

An immediate observation reveals that most CIs cross zero, reflecting a high level of uncertainty in the estimates. This outcome is expected, given the reduced number of data points within each subset, consequently leading to larger standard errors and broader CIs relative to the aggregated effects (with all categories combined). Among the six CSP categories, *CC-V Linking* emerges as the category showing the highest and most robust overall estimate (Overall: b = 0.32) with the CIs not crossing zero. Although *Deletion of Onset* (Overall: b = 0.22), *Reduction of Stop* (Overall: b = 0.23), and *Palatalisation* (Overall: b = 0.22) also trend towards overall improvement, their corresponding CIs all include zero, indicating statistical uncertainty. In addition, within each category, the CIs for each of the conditions overlap considerably, suggesting no evidence of differential improvement between groups. We note that the ER group in *Deletion of Onset* (ER: b = 0.59) and the EN group in *CC-V Linking* (EN: b = 0.55) exhibit CIs that do not cross zero. However, while these indicate potential group-specific improvement, again the substantially overlapping CIs within these categories suggest that no statistically robust differences between groups can be concluded.

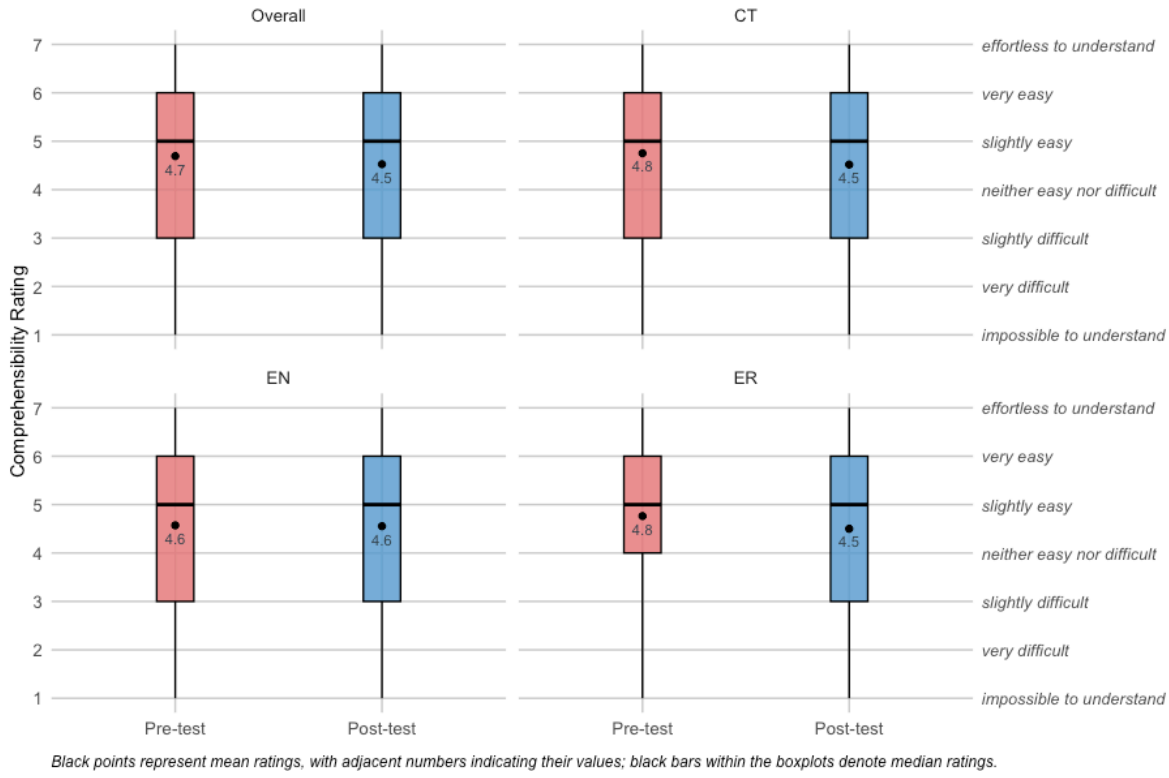
5.2.4 Comprehensibility Rating

Figure 5.26 illustrates participants' comprehensibility ratings for connected speech items across the two assessment time points. Notable similarities can be observed across groups, with median ratings consistently at 5 (*slightly easy to understand*). The interquartile range spans four rating levels – from 3 (*slightly difficult to understand*) to 6 (*very easy to understand*). The only exception is the pre-test data for the ER group, which exhibits a

slightly narrower interquartile range spanning three rating levels from 4 (*neither easy nor difficult*) to 6. Overall, mean ratings across groups and assessments all exceed 4.5, indicating that participants generally did not perceive the connected speech items as particularly challenging to understand.

Figure 5.26

Comprehensibility Ratings by Group and Assessment: Boxplot Distributions



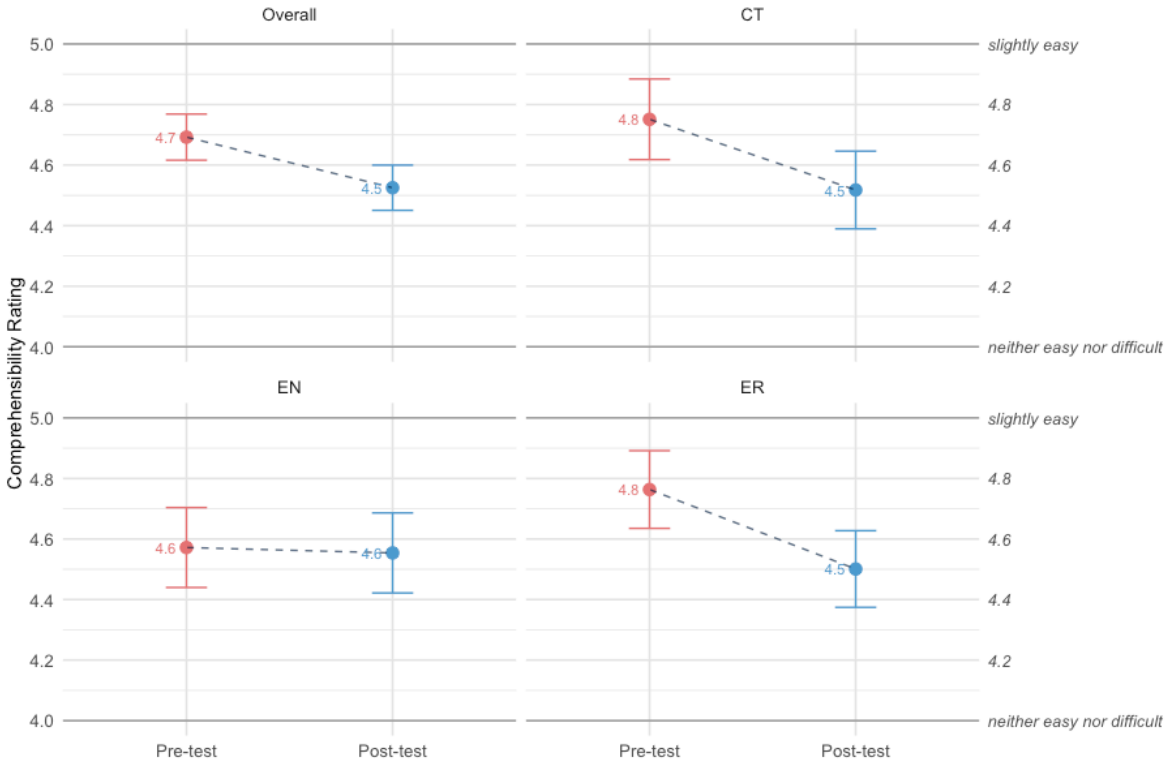
Black points represent mean ratings, with adjacent numbers indicating their values; black bars within the boxplots denote median ratings.

Despite ratings consistently leaning towards the 'easy' end of the scale, an unexpected trend emerges: comprehensibility ratings slightly decline from pre-test (overall mean = 4.7) to post-test (overall mean = 4.5), suggesting that participants overall perceived connected speech as *less* comprehensible following the interventions. This decrease is numerically more pronounced in the ER and CT groups (each showing a reduction of 0.3 points), whereas the EN group displays no numerical change. This trend is further

visualised in Figure 5.27, with mean ratings and 95% confidence intervals. Bayesian inferential analyses were subsequently conducted to examine these patterns.

Figure 5.27

Comprehensibility Ratings by Group and Assessment: Means with Confidence Intervals



Bayesian Ordinal Mixed Effects Regression Model

To analyse these ordinal comprehensibility ratings (from 1 = *impossible to understand* to 7 = *effortless to understand*), a Bayesian ordinal mixed effects regression model was fitted using the *brms* package. Two versions of the model, consistent with previous analyses (cf. Sections 5.2.1 and 5.2.2), were employed to evaluate whether the decline in comprehensibility ratings was statistically robust. Table 5.10 summarises the model output for the intercepts and key fixed effects parameters.

Model for Comprehensibility Ratings: Version 1

$$Comprehensibility\ rating \sim CT\_VERSUS\_ER * Assessment.ct + CT\_VERSUS\_EN * Assessment.ct + LexTALE.ct + (Assessment.ct \mid PID)$$

Model for Comprehensibility Ratings: Version 2

$$Comprehensibility\ rating \sim ER\_VERSUS\_CT * Assessment.ct + ER\_VERSUS\_EN * Assessment.ct + LexTALE.ct + (Assessment.ct \mid PID)$$

Table 5.10

Summary of Bayesian Model Output for Comprehensibility Ratings

	Estimate	**Est. Error**	**CI_lower**	**CI_upper**	**B<>0**
Intercept[1]	-3.83	0.14	-4.11	-3.55	1
Intercept[2]	-2.27	0.11	-2.46	-2.04	1
Intercept[3]	-1.15	0.11	-1.37	-0.93	1
Intercept[4]	-0.41	0.11	-0.63	-0.19	1
Intercept[5]	0.74	0.11	0.52	0.96	1
Intercept[6]	2.06	0.11	1.84	2.29	1
LexTALE.ct	1.06	0.87	-0.11	3.26	0.97
Assessment.ct	-0.24	0.10	-0.44	-0.03	0.99
Assessment.ct (CT)	-0.35	0.23	-0.80	0.11	0.94
Assessment.ct (ER)	-0.37	0.16	-0.69	-0.04	0.99
Assessment.ct (EN)	-0.02	0.16	-0.34	0.30	0.55
CT_VERSUS_ER:Assessment.ct	0.04	0.26	-0.46	0.54	0.56

CT_VERSUS_EN:Assessment.ct	0.34	0.25	-0.14	0.84	0.92
ER_VERSUS_EN:Assessment.ct	0.31	0.26	-0.18	0.83	0.88

Note that the ordinal regression model estimates multiple intercepts corresponding to rating thresholds, each representing cumulative log-odds of a response falling at or below a given rating. For example, Intercept[1] (b = -3.83) represents the cumulative log-odds of receiving a rating of 1 (as there is no rating below 1). These intercepts can be converted into cumulative probabilities, from which the probabilities for individual ratings can be mathematically derived (Table 5.11).

Table 5.11

Bayesian Model Intercepts and Corresponding Probabilities for Individual Comprehensibility Rating Categories

Intercept	**Rating Category**	**Estimate (Log-Odds)**	**Cumulative Probability**	**Category Probability**
Intercept[1]	1 (*impossible to understand*)	-3.83	2.12%	2.12%
Intercept[2]	2 (*very difficult*)	-2.27	9.36%	7.24%
Intercept[3]	3 (*slightly difficult*)	-1.15	24.05%	14.69%
Intercept[4]	4 (*neither easy nor difficult*)	-0.41	39.89%	15.84%
Intercept[5]	5 (*slightly easy*)	0.74	67.70%	27.81%
Intercept[6]	6 (*very easy*)	2.06	88.70%	21.00%
-	7 (*effortless to understand*)	(∞)	(100.00%)	(11.30%)

The probability for each rating is computed by subtracting the cumulative probability of a preceding rating from that of the target rating. Using the six intercept estimates provided

in the model output, the probability of the highest rating of 7 can be logically derived (values shown in parentheses). In line with the general patterns illustrated in Figure 5.26, the most probable rating is 5 (27.81%), followed by ratings of 6 (21.00%) and 4 (15.84%). Ratings below the neutral point of 4 generally display lower probabilities compared to those above it, from a rating of 3 (14.69%), to ratings of 2 (7.24%) and 1 (2.12%).

Unlike in previous measures (connected speech accuracy and disparity scores), the CIs for LexTALE.ct (b = 1.06, CIs[-0.11, 3.26]) cross zero, indicating uncertainty regarding the effect of proficiency on comprehensibility ratings. Crucially, the main effect of Assessment.ct exhibits a negative estimate (b = -0.24, CIs[-0.44, -0.03]) and the CIs do not cross zero, providing evidence of a decline in comprehensibility ratings from the pre-test to post-test.

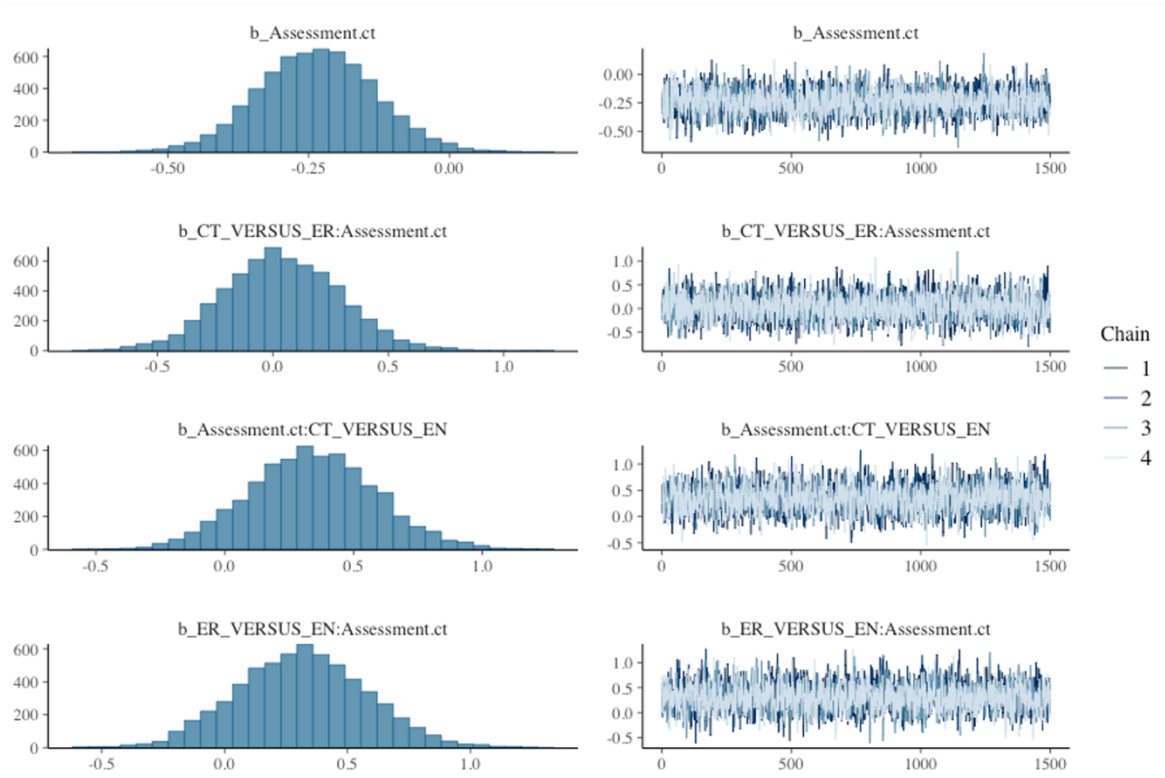
Group-specific effects indicate that the ER group shows the clearest decline (b = -0.37, CIs[-0.69, -0.04]), with the CIs entirely below zero. In contrast, the CIs for the CT group (b = -0.35, CIs[-0.80, 0.11]) and the EN group (b = -0.02, CIs[-0.34, 0.30]) cross zero, suggesting uncertainty about these declines. Additionally, all interaction terms have CIs crossing zero (CT_VERSUS_EN:Assessment.ct: b = 0.34, CIs[-0.14, 0.84]; ER_VERSUS_EN:Assessment.ct: b = 0.31, CIs[-0.18, 0.83]; CT_VERSUS_ER:Assessment.ct: b = 0.04, CIs[-0.46, 0.54]), providing no evidence of meaningful differences between groups regarding the magnitude of changes in comprehensibility ratings. Therefore, similar to the disparity score metric, while individual group effects show differing trends, no statistically robust interactions can be identified.

Figure 5.28 presents the posterior density distributions and MCMC sampling trace plots

for the main effect and interaction parameters. Corresponding to the B<>0 values in Table 5.10, the vast majority (99%) of sampling iterations for the main effect of Assessment.ct yield negative estimates, indicating a strong trend towards decreased ratings after the interventions. High proportions of positive estimates for CT_VERSUS_EN:Assessment.ct (92%) and ER_VERSUS_EN:Assessment.ct (88%) suggest high probabilities that the EN group experienced smaller declines in ratings (reflecting its identical means at the pre-test and post-test, cf. Figure 5.27) compared to the CT and ER groups, though these differences remain statistically uncertain. In contrast, the posterior distribution for CT_VERSUS_ER:Assessment.ct is approximately evenly split (with 56% of iterations being positive and 44% negative), reflecting considerable uncertainty regarding the direction of the interaction effect for this group contrast across assessments.

Figure 5.28

Bayesian Posterior Density Distributions for Key Parameters in Comprehensibility Rating Model



Bayes Factor Analysis

As with disparity scores, we conducted Bayes factor analysis to explore whether the data supports the null for the interaction effects in comprehensibility ratings. Following the procedures employed previously (cf. Section 5.2.2: *Bayes Factor Analysis*), Dienes's (2008) BF calculator (see Appendix H for the R code) was used, with the required sample information (i.e. parameter estimates and standard errors) extracted from Table 5.10. Predicted effects for each interaction were derived using the same motivated-maximum approach (Silvey et al., 2024), wherein an interaction was based on the average effect of assessment (pre-test vs. post-test) across the two groups being compared (see Appendix I). Three separate models, each examining a pairwise group contrast with the same random effect structure, provided predicted interaction effects. The relevant estimates for the average effects of assessment for each of the group contrasts were:

(a) CT and ER groups: -0.36

(b) CT and EN groups: -0.18

(c) ER and EN groups: -0.16

The absolute values of these estimates were set as the standard deviations for modelling H_1 of the corresponding interaction terms (cf. Section 5.2.2: *Bayes Factor Analysis* for the rationale of this specification).

Table 5.12 presents the inputs and results of the BF computations for the target interaction terms. Robustness regions are included to show the range of predicted effects that would yield a BF within the same evidential category (i.e. between 1/3 and 3 in each case). Note that since the absolute values of the predicted effects were used as the standard deviations of H_1 with a mode of 0, and that these BF tests were two-tailed, the sign of these predicted effects had no influence on the BF results.

Table 5.12

Summary of Bayes Factors for Interaction Parameters for Comprehensibility Ratings

	Sample Estimate	**Sample SE**	**Predicted Effect**	**Bayes Factor**	**Robustness Region**
CT_VERSUS_ER:Assessment.ct	0.04	0.26	0.36	0.59	[-0.75, 0.75]
CT_VERSUS_EN:Assessment.ct	0.34	0.25	0.18	1.11	[-1.86, 1.86]
ER_VERSUS_EN:Assessment.ct	0.31	0.26	0.16	1.04	[-1.55, 1.55]

Note. The first three columns represent inputs to the calculator. The final column provides the range of values of predicted effects which would produce the same qualitative result (i.e. a BF between 1/3 and 3 indicating ambiguous evidence).

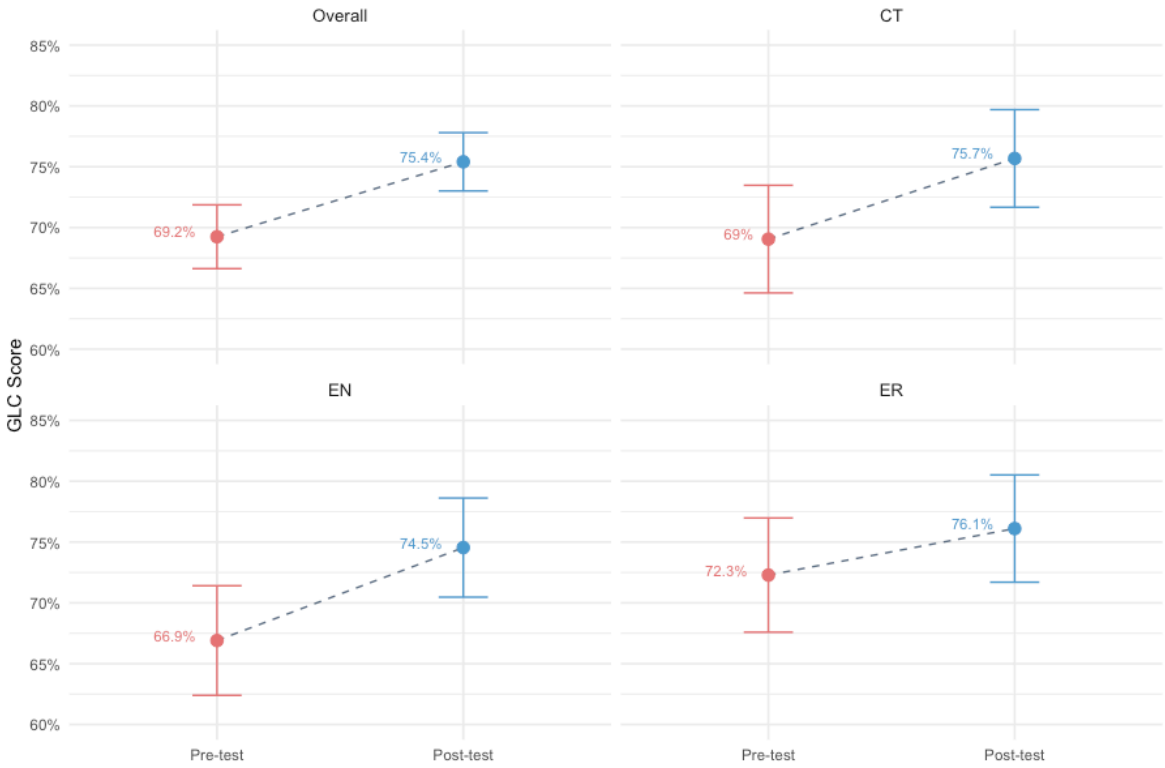
Following Jeffrey's (1961) interpretative guidelines (Table 5.9), all computed BFs fall within the ambiguous region (between 1/3 and 3). Specifically, CT_VERSUS_ER:Assessment.ct ($BF_{(0, 0.36)}$ = 0.59) leans slightly towards H_0, whereas CT_VERSUS_EN:Assessment.ct ($BF_{(0, 0.18)}$ = 1.11) and ER_VERSUS_EN:Assessment.ct ($BF_{(0, 0.16)}$ = 1.04) indicate almost equal likelihood for both H_1 and H_0. Crucially, the ambiguity of these BFs suggests that no conclusions can be drawn regarding the null effects for the interaction terms observed in the previous Bayesian model output (Table 5.10) – i.e. it remains unclear whether H_0 is plausible.

5.2.5 L2 General Listening Comprehension (GLC)

The GLC scores from the pre-test and post-test were analysed to determine whether the interventions had a broader effect on participants' general listening comprehension. Importantly, the GLC assessment materials were also counterbalanced across the pre-test and post-test to mitigate potential confounding influences, such as differences in topic familiarity or intrinsic item difficulty. Figure 5.29 presents the mean GLC scores by group and assessment along with 95% confidence intervals.

Figure 5.29

General Listening Comprehension Scores by Group and Assessment



At the pre-test, the ER group demonstrates the highest mean score (72.3%), followed by the CT group (69.0%) and the EN group (66.9%). After the interventions, all groups display an upward trend, with their relative rankings remaining unchanged. Specifically, the ER group increases by 3.8 percentage points to a post-test mean of 76.1%, the CT group increases by 6.7 percentage points to 75.7%, and the EN group increases by 7.6 percentage points to 74.5%. Aggregated across all groups, the overall mean GLC score improve by 6.2 percentage points from the pre-test (69.2%) to post-test (75.4%).

Bayesian Logistic Mixed Effects Regression Model

To examine the certainty of the observed improvement and group differences across assessments, a Bayesian logistic mixed effects regression model was fitted (with the two complementary versions including all pairwise contrasts). The model included the same

random effects (Assessment.ct | PID) as specified previously for the principal measures. Table 5.13 summarises the model output for the key fixed effects parameters.

Model for GLC: Version 1

$$GLC\ score \sim CT\_VERSUS\_ER * Assessment.ct + CT\_VERSUS\_EN * Assessment.ct + (Assessment.ct \mid PID)$$

Model for GLC: Version 2

$$GLC\ score \sim ER\_VERSUS\_CT * Assessment.ct + ER\_VERSUS\_EN * Assessment.ct + (Assessment.ct \mid PID)$$

Table 5.13

Summary of Bayesian Model Output for General Listening Comprehension

	Estimate	**Est. Error**	**CI_lower**	**CI_upper**	**B<>0**
Intercept	1.06	0.09	0.89	1.24	1
Assessment.ct	0.25	0.14	-0.01	0.52	0.96
Assessment.ct (CT)	0.32	0.18	-0.04	0.67	0.96
Assessment.ct (ER)	0.11	0.33	-0.56	0.74	0.64
Assessment.ct (EN)	0.34	0.26	-0.18	0.85	0.90
CT_VERSUS_ER:Assessment.ct	-0.18	0.34	-0.86	0.50	0.70
CT_VERSUS_EN:Assessment.ct	0.03	0.32	-0.60	0.66	0.52
ER_VERSUS_EN:Assessment.ct	0.21	0.35	-0.46	0.92	0.73

Note. Estimates are in log-odds space.

The main effect of Assessment.ct (b = 0.25, CIs[-0.01, 0.52]) indicates an overall trend towards improvement in GLC scores from the pre-test to post-test, while the CIs

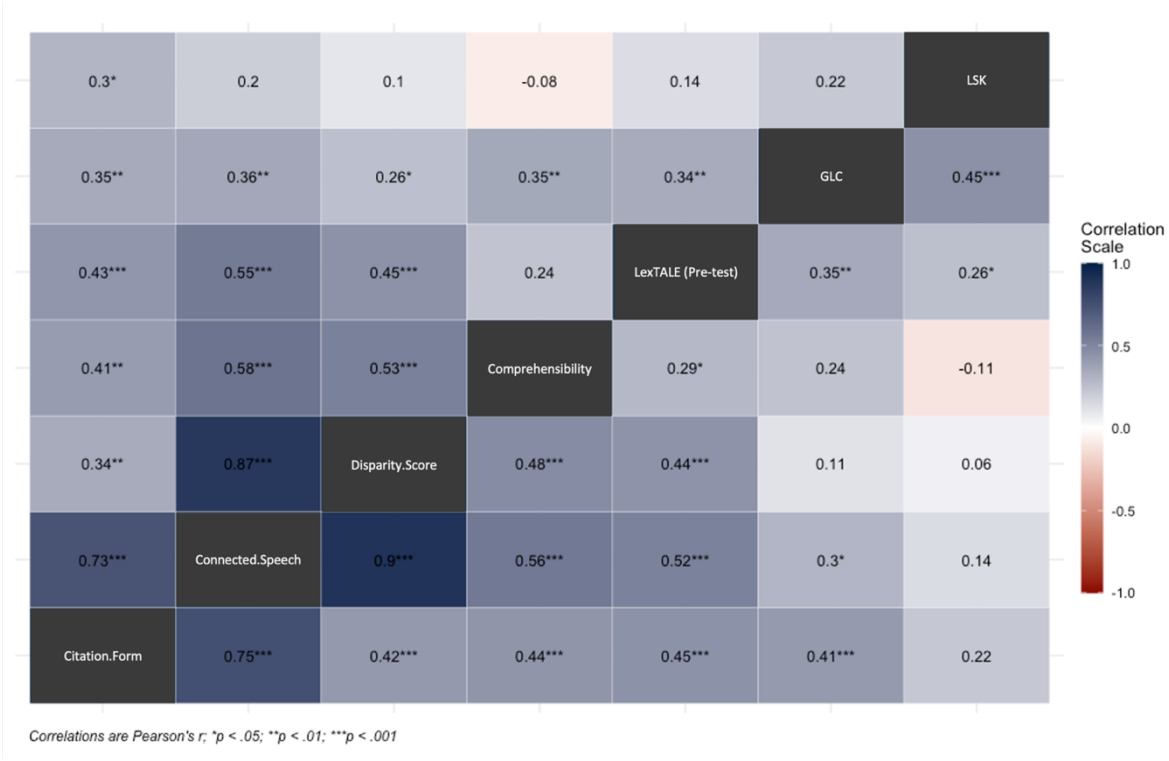
marginally cross zero, reflecting minor statistical uncertainty. The estimates for individual groups are similarly accompanied by CIs that cross zero (CT: b = 0.32, CIs[-0.04, 0.67]; ER: b = 0.11, CIs[-0.56, 0.74]; EN: b = 0.34, CIs[-0.18, 0.85]), though the CT group shows relatively stronger yet still uncertain evidence of improvement. Moreover, all interaction terms demonstrate great uncertainty (CT_VERSUS_ER:Assessment.ct: b = -0.18, CIs[-0.86, 0.50]; CT_VERSUS_EN:Assessment.ct: b = 0.03, CIs[-0.60, 0.66]; ER_VERSUS_EN:Assessment.ct: b = 0.21, CIs[-0.46, 0.92]), suggesting no statistically robust evidence of differential improvement between groups.

5.2.6 Correlations among Linguistic Measures

To explore the interrelationships among the linguistic measures employed in this study – including the disparity score metric, which specifically indexes CSP-induced difficulty (see Section 5.2.2) – a correlation analysis was conducted. Figure 5.30 presents a correlation matrix encompassing all the measures at both pre-test and post-test. The lower-right section of the diagonal shows pre-test correlations, and the mirrored upper-left section shows post-test correlations, allowing for direct comparisons across the two assessment time points. Gradient colours illustrate the magnitude and direction of correlations, with blue indicating positive and red indicating negative correlations.

Figure 5.30

Correlation Matrix for Linguistic Measures: Pre-test versus Post-test



Note. The lower-right section represents pre-test correlations; the upper-left section represents post-test correlations.

Principal Measures

Correlations are generally comparable across the pre-test and post-test, particularly among the principal measures (the lower-left 4x4 matrix in Figure 5.30). Strong positive correlations are observed between connected speech and citation form (pre-test: $r = 0.75$; post-test: $r = 0.73$). The disparity score, as expected, exhibits stronger correlations with connected speech (pre-test: $r = 0.90$; post-test: $r = 0.87$) than with citation form (pre-test: $r = 0.42$; post-test $r = 0.34$). This is because the disparity score reflects the degree of CSP-induced difficulty (reverse coded so that higher values indicate stronger performance – i.e. less difficulty), which occurs only in connected speech, and also because connected speech scores show greater variability than citation form scores (cf. Figure 5.19).

In this study, these three accuracy-based metrics are operationalised as three distinct forms of intelligibility – citation form intelligibility, connected speech intelligibility, and relative intelligibility (as captured by the disparity score). This operationalisation enables an exploration of how these accuracy-based intelligibility metrics relate to participants' comprehensibility ratings, which represent their perceived level of ease or difficulty in understanding stimuli containing the target CSPs. As shown in Figure 5.30, comprehensibility demonstrates moderate correlations with citation form (pre-test: r = 0.44; post-test: r = 0.41), connected speech (pre-test: r = 0.56; post-test: r = 0.58), and disparity score (reverse coded; pre-test, r = 0.48; post-test: r = 0.53).

Further, to assess whether the correlations between comprehensibility and all three intelligibility metrics differ significantly at each assessment time point, a *Z*-test was conducted using Fisher's (1921) *r*-to-*z* transformation, also known as the *atanh* function. This transformation essentially converts a relatively skewed *r* distribution with bounded intervals [-1, 1] into a more normally distributed *z*-distribution with unbounded intervals [-∞, ∞] suitable for inferential testing[35]. Specifically, for each correlation coefficient (r), the transformed value (Zr) was computed using Fisher's formula:

$$Zr = \text{atanh}\,(r) = \frac{1}{2}\log\frac{1+r}{1-r}$$

Subsequently, these Zr values were used for the calculation of *Z*-statistic, following the approach outlined in Cohen et al. (2002):

[35]A theoretical discussion can be found in Cox (2008).

$$Z = \frac{Zr_1 - Zr_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

Table 5.14 summarises the *Z*-test results. An absolute *Z*-value exceeding 1.96 corresponds to a *p*-value below .05, indicating a statistically significant difference between the correlations compared.

Table 5.14

Z-test Results for Correlations between Comprehensibility and Three Intelligibility Metrics

Pairwise Comparisons		**Pre-test**			**Post-test**		
Correlation 1	**Correlation 2**	***Zr₁***	***Zr₂***	***Z*-value**	***Zr₁***	***Zr₂***	***Z*-value**
Comprehensibility/ Citation Form	Comprehensibility/ Connected Speech	0.47	0.63	-0.89	0.44	0.66	-1.23
Comprehensibility/ Citation Form	Comprehensibility/ Disparity Score	0.47	0.52	-0.28	0.44	0.59	-0.84
Comprehensibility/ Connected Speech	Comprehensibility/ Disparity Score	0.63	0.52	0.61	0.66	0.59	0.39

As shown in Table 5.14, none of the pairwise comparisons display a *Z*-value exceeding the critical threshold of ±1.96 (corresponding to *p*-value < .05), indicating that there is no evidence that comprehensibility ratings correlate significantly more strongly with any one intelligibility metric compared to another at either assessment time point. This suggests that participants' subjective perception of ease in understanding connected speech is generally associated to a similar extent with citation form intelligibility, connected speech intelligibility, and relative intelligibility (i.e. the disparity score), despite differing patterns

within these individual measures (as detailed in Sections 5.2.1 and 5.2.2 above).

Supplementary Measures

Correlations involving supplementary measures generally show weak to moderate coefficients. The LexTALE (administered at the pre-test as a proxy for proficiency) correlates moderately with all principal measures, except with comprehensibility (pre-test: $r = 0.29$, post-test: $r = 0.24$). Similarly, despite greater variability, the GLC correlates moderately with most other measures, with exceptions being its weaker pre-test correlations with disparity score ($r = 0.11$) and comprehensibility ($r = 0.24$). On the other hand, the LSK largely yields weak, non-significant correlations with other measures as expected, due to the deliberate control over vocabulary level (cf. Section 3.3.2: *L1 Lexical-Semantic Knowledge Test* (LSK)); however, an unexpected moderate, significant correlation is observed between the LSK and GLC at the pre-test ($r = 0.45$).

To determine whether these correlations significantly changed from the pre-test to post-test, the same *Z*-statistic approach demonstrated above was applied to individual pairwise pre-test and post-test coefficients depicted in Figure 5.30. Specifically, for each pair of correlations, two *Zr* values were computed independently: one for the pre-test (denoted as $Zr_{pre\text{-}test}$) and one for the post-test (denoted as $Zr_{post\text{-}test}$). These values were then used to calculate the corresponding *Z*-statistic. These results are summarised in Table 5.15.

Table 5.15

Z-test Results for Pairwise Comparisons of Pre-test versus Post-test Correlations

Measure 1	**Measure 2**	$Zr_{pre\text{-}test}$	$Zr_{post\text{-}test}$	***Z*-value**
Citation Form	Connected Speech	0.97	0.93	0.25

Citation Form	Disparity Score	0.45	0.35	0.53
Citation Form	Comprehensibility	0.47	0.44	0.20
Citation Form	LexTALE (Pre-test)	0.48	0.46	0.14
Citation Form	GLC	0.44	0.37	0.38
Citation Form	LSK	0.22	0.31	-0.48
Connected Speech	Disparity Score	1.47	1.33	0.78
Connected Speech	Comprehensibility	0.63	0.66	-0.16
Connected Speech	LexTALE (Pre-test)	0.58	0.62	-0.24
Connected Speech	GLC	0.31	0.38	-0.36
Connected Speech	LSK	0.14	0.20	-0.34
Disparity Score	Comprehensibility	0.52	0.59	-0.37
Disparity Score	LexTALE (Pre-test)	0.47	0.48	-0.07
Disparity Score	GLC	0.11	0.27	-0.84
Disparity Score	LSK	0.06	0.10	-0.22
Comprehensibility	LexTALE (Pre-test)	0.30	0.24	0.29
Comprehensibility	GLC	0.24	0.37	-0.64
Comprehensibility	LSK	-0.11	-0.08	-0.16
LexTALE (Pre-test)	GLC	0.37	0.35	0.06
LexTALE (Pre-test)	LSK	0.27	0.14	0.69
GLC	LSK	0.48	0.22	1.39

It can be seen that none of the pairwise comparisons yields a *Z*-value exceeding the critical threshold of ±1.96 (corresponding to $p < .05$), indicating no evidence of statistically significant changes in correlations among the linguistic measures from the pre-test to post-test. Thus, the overall structure of interrelationships among the linguistic measures

remains stable over time.

5.3 Summary of Phase 2 Study

Phase 2 investigated the effectiveness of pedagogical interventions targeting the six CSP categories identified as particularly challenging in Phase 1. A total of 66 university students in Taiwan, possessing proficiency at the CEFR B2 level or above, participated in the study. Participants were randomly assigned to one of three groups, each receiving a distinct pedagogical intervention: a rule-based condition, a noticing-based condition, and an implicit condition (serving as a control group). This study employed a pre-test-intervention-post-test design, utilising the same linguistic measures from Phase 1, with the addition of a comprehensibility rating task. The stimuli were selected from the inventory developed and validated in Phase 1 and were counterbalanced across the pre-test and post-test. The language input provided during the interventions was meticulously controlled, ensuring that participants in all groups encountered identical targeted CSP sample materials, notwithstanding differences in pedagogical approaches.

Data were analysed using Bayesian mixed effects regression models supplemented by Bayes factors where appropriate. The primary findings indicated that overall participants' transcription accuracy in connected speech improved following the interventions, accompanied by a reduction in CSP-induced difficulties (as measured by disparity scores). While the experimental groups tended to show numerically greater improvement – largely driving the main effects – the control group exhibited comparatively minimal changes. However, the Bayesian models revealed substantial statistical uncertainty regarding the group-by-assessment interaction effects. Further Bayes factors analyses yielded ambiguous evidence in favour of neither the presence (H_1) nor the absence (H_0) of group

differences. Similar ambiguity emerged in the examination of intervention effects on individual CSP categories, showing no evidence of group differences. Interestingly, comprehensibility ratings displayed an unexpected inverse pattern, where overall participants' subjective perception of ease in understanding connected speech declined following the interventions, despite improvement in transcription-based intelligibility measures. Nevertheless, there was again no evidence that this decline differed by intervention condition.

In addition, analyses of the supplementary measure of general listening comprehension (GLC) indicated no evidence of statistically robust improvement from the pre-test to post-test, nor was there evidence supporting differential effects across the intervention conditions. Correlation analyses highlighted positive, significant associations among the principal measures. Subsequent *Z*-tests showed no significant differences in the magnitude of correlations between comprehensibility ratings and each of the three intelligibility metrics (i.e. connected speech, citation form, and disparity score) at either assessment time point. Similarly, the supplementary measures of LexTALE and GLC were found to positively correlate with both connected speech and citation form, whereas the LSK yielded mostly weak and non-significant correlations with other measures. Pairwise *Z*-tests further demonstrated consistency in the overall correlation structure, with no significant changes in any correlations from the pre-test to post-test.

Chapter 6: Discussion

The Phase 1 (diagnosis) and Phase 2 (intervention) studies have in turn (a) identified the difficulties that adult Mandarin-speaking learners of English face in processing speech containing CSPs and (b) evaluated the effectiveness of three distinct pedagogical interventions – two experimental conditions targeting particularly challenging CSP categories in English, and an implicit control condition. This chapter discusses the findings from both phases, exploring their theoretical and practical implications. First, I discuss potential accounts for the varied patterns observed across different CSP categories, addressing RQ 1a and 1b. Next, I probe the effectiveness of the interventions, addressing RQ 2a to RQ 2e. Finally, I explore some practical considerations for teaching CSPs.

6.1 RQ 1a & 1b: Varied Impact of CSPs: Phonological Mechanisms and Misperceptions

As a reminder, Research Questions 1a and 1b were as follows:

RQ 1. How do English CSPs affect adult Mandarin-speaking learners' L2 speech perception?

RQ 1a. Does the L2 learners' ability to recognise CSP-affected words vary across different categories of CSPs?

RQ 1b. If so, which categories of CSPs are particularly challenging for the L2 learners?

The diagnostic data reveal considerable variability in the difficulties associated with different categories of CSPs. This is evidenced in the transcription accuracy disparities between citation form and connected speech, which range from 13 to 49 percentage points across the ten CSP categories investigated (cf. Figure 4.8). Since each CSP is

characterised by its unique phonological mechanisms and, in certain cases, relevant lexical contexts, they are discussed individually below. The following sections specifically focus on the six CSP categories identified as most challenging in Phase 1, which were subsequently targeted in Phase 2. Note that to obtain a richer illustration of linguistic nuances associated with target CSP categories, the discussion here draws on representative instances of misperception from both Phase 1 and the pre- and post-tests in Phase 2. Misperception patterns were largely consistent across these, though Phase 2 provides more error instances than Phase 1, due to the larger and slightly less proficient participant group (cf. Sections 4.1.1 and 5.1.1). Both pre-test and post-test data were considered in Phase 2, as the stimuli were counterbalanced, and comparable errors appeared across both assessment time points. L1 listener data serve as a baseline for comparison, which is addressed following the discussion of each CSP category.

Deletion of Onset

Deletion of Onset, which refers to the complete loss of an initial segment (e.g. *watch him* becoming /wɒt͡ʃɪm/), was identified as the most challenging category among the investigated CSPs with the greatest accuracy disparity of 49 percentage points between citation form and connected speech. For L2 listeners, this substantial difficulty likely stems from the drastic phonological change involved – the complete loss of a sound. Since the deletion occurs at the onset position, it may disrupt the initial activation of acoustic-phonetically plausible items in the mental lexicon, potentially eliminating the correct target from the candidate pool at an early stage (Marslen-Wilson & Welsh, 1978). Additionally, the remaining (undeleted) part of the word is typically a single unstressed syllable containing a lax vowel, as this CSP mainly affects function words such as pronominal forms and auxiliaries. These factors can increase the likelihood of

misperception, with the target item often being erroneously identified as other monosyllabic function words without a breathed onset (/h/). For example, in our data, the deletion of /h/ in *watch him* may lead to its misperception as *watch it*, or *that's hers* as *that's us*, wherein the lax vowels following the onset deletion resemble those of the misperceived items. Notably, while these misperceptions may exhibit phonological similarity to the intended targets, they often fail to fit syntactically or semantically into the given sentential contexts[36] (this is also true of most examples discussed in other categories below). This phenomenon suggests that, while top-down and bottom-up processing may be assumed to function in an interactive manner (e.g. Magnuson, 2017), L2 listeners in these cases appear to rely more heavily on their auditory perception than on broader syntactic or pragmatic cues.

Glide Insertion

Glide Insertion involves the addition of a glide (/j/ or /w/) between two vowels across word boundaries (e.g. *blue ink* realised as /bluːwɪŋk/). This category emerged as the second most challenging category, with an accuracy disparity of 33 percentage points between citation form and connected speech. While all segments present in the citation form are retained, an inserted glide can substantially alter the perceived phonological structure. Specifically, as the glottal stop that typically precedes the articulation of a word-initial vowel in citation form is 'smoothed out' by a glide (which is often referred to as a semivowel due to its similar formant characteristics to certain vowels[37]), the perception of word boundaries can be compromised. Furthermore, the L2 learners may also be influenced by the 'canonical' orthographic representations of /w/ as <w> and /j/ as <y>,

[36]See Appendix C for a full list of stimuli and corresponding sentential contexts.

[37]Refer to Ladefoged and Johnson (2015) for a detailed discussion.

likely as a result of the strong sound-spelling correspondences in their mind. A recurring example in the data involved the misperception of *blue ink* (/bluːwɪŋk/) as **blewing*, wherein the inserted glide /w/ between the two words was reflected as part of a homophone of the first word (*blew*), represented by the letter <w> and suffixed by the inflectional morpheme *-ing*, rather than being recognised as a transition to the second word *ink*. This case illustrates that the inserted glide can alter word segmentation, potentially prompting the listener to first select a homophonous candidate that exhibits stronger orthographic correspondence with the auditory perception of the inserted glide, and then to attach the inflectional morpheme *-ing* to *blew* based on its morphological viability (as a verb conjugation) and phonological similarity, despite the resultant form (**blewing*) being ungrammatical in the context.

Reduction of Stop

Reduction of Stop occurs when the release burst of a word-final stop is reduced before another stop or an affricate (e.g. *rob Jack* realised as /rɒbˀ.dʒæk/). This category ranked third in difficulty, with an accuracy disparity of 29 percentage points between citation form and connected speech. The reduced release can render the whole segment perceptually ambiguous, because the place of articulation of the stop may only be fully audible through the air release. Erroneous responses in the data indicated that reduced stops were frequently misperceived as other stops that would have been subject to the same reduction process. For example, /b/ in *rob Jack* was often misperceived as /p/, yielding *rope Jack,* or as /d/, yielding *road Jack*. Although vowel quality (/ɔ/ in *rob* vs. /əʊ/ in *rope* and *road*) and vowel length (determined by the following voiced /b/ vs. voiceless /p/) may theoretically provide distinguishable phonological cues, these cues appeared insufficient to ‘rectify’ such misperceptions. Additionally, when a word-final

reduced stop is followed by an unreduced stop at the onset of the subsequent word, the latter can sometimes be misperceived as replacing the former. For example, frequent misperceptions for *pet cat* included *pack*, *packed*, *packet*, or *package*, suggesting that the reduced coda /t/ in *pet* was substituted with the fully articulated onset /k/ in *cat*, with the remaining segments either disregarded or merged as part of a single concatenated word rather than the intended two-word sequence. These cases illustrate how unreleased stops can affect L2 listeners' ability to parse segments into words, particularly when two stops occur in succession, with one undergoing non-release and the other fully articulated.

CC-V Linking

*CC-V Linkin*g involves resyllabification, wherein the final consonant of a word is realised as the onset of the following syllable across word boundaries (e.g. *wasp or* realised as /wɒs.pɔː/). It was identified as the most challenging subcategory of linking, with an accuracy disparity of 26 percentage points between citation form and connected speech. L2 listeners' difficulty with this CSP was evident in misperceptions such as *Welsh office* misperceived as *well shelf*, and *wasp or* misperceived as **wasble* (with variations including **wasbal* or **wusbble*). In both cases, the last segment of the word-final CC cluster (/ʃ/ and /p/, respectively) was misperceived as the onset of the following syllable, aligned with their realisation in connected speech.

Notably, some frequent erroneous responses linked to this CSP appear to have been driven by pragmatic cues, such as *bulb at* in the item '*It's a glass tube with a <u>bulb at</u> the bottom*' being misperceived as *bowl at*, and *bookshelf invites* in '*Their <u>bookshelf invites</u> a lot of readers*' being misperceived as *bookshop invites*. In both cases, the second lexical item was correctly identified, while the first item was not. Crucially, in addition to their

phonological similarity to the intended targets, the misperceived words remained syntactically permissible within the given contexts. On the one hand, these errors validate the strategic design and piloting of the stimuli, which aimed to control for predictability and prevent listeners from inferring the answers without fully processing the auditory input (see Section 4.1.3 for the devising and piloting of stimuli). On the other hand, while these errors may not be explicitly attributable to the process of resyllabification, the response patterns possibly suggest an indirect impact of the CSP. Specifically, in cases where no plausible candidate aligns with the auditory perception of the resyllabified target, the listeners may default to a phonologically similar item that best fits the syntactic and pragmatic context. This compensatory phenomenon highlights the role of top-down processing, particularly in contrast to non-word and ungrammatical responses discussed above in this and other categories. Meanwhile, the varied patterns in response to different stimuli underscore the complexity of speech perception and the dynamic interplay between top-down and bottom-up processing (Vandergrift, 2004).

Palatalisation

Palatalisation involves the coalescence of an alveolar segment and a following /j/ into a post-alveolar segment (e.g. *makes you* realised as /meɪkʃuː/). Among the two subcategories of modification examined, *Palatalisation* exhibited a larger accuracy disparity of 25 percentage points between citation form and connected speech. The impact of a palatalised segment was most evident in cases that involved alveolar fricatives /s/ and /z/, such as *makes you* being misperceived as *make she*, *that's your* as *that short*, and *as you* as *as she*. In these examples, the post-alveolar fricative /ʃ/ was misperceived as the onset of the second word, rather than as the outcome of palatalisation that involved two segments across word boundaries, thereby affecting the listener's lexical selection.

Notably, the word *she* may have been a particularly tempting choice due to its shared lexical category (i.e. pronoun) with the intended target *you*, which is frequently affected by this CSP (Celce-Murcia et al., 2010). This lexical resemblance also made some erroneous responses both syntactically and pragmatically permissible – for example, *as she* replacing the intended targets *as you* in the item '*The train was cancelled again, as you said'*. While such cases are less likely to cause immediate comprehension failure compared to other pragmatically nonsensical misperceptions, their contextual plausibility could actually increase the risk of undetected miscomprehension, as the listener might not immediately notice the substitution until it becomes apparent and requires clarification later in the discourse.

Regressive Assimilation

Finally*, Regressive Assimilation* involves the alveolar nasal /n/ being assimilated to the bilabial /m/ or velar /ŋ/ (e.g. *on point* realised as /ɒmpɔɪnt/). In this category, the phonological structure across word boundaries remains intact, and only the place of articulation is modified, while the nasal manner is preserved. This CSP has received limited attention in the literature, perhaps because it is regarded as comparatively less salient. However, despite the relatively minor alteration, the current study suggests that it appears sufficiently perceptible to influence L2 speech perception. Notably, the diagnostic data revealed an accuracy disparity of 20 percentage points between citation form and connected speech. Examples of erroneous responses included *on point* being misperceived as *important* or *unimportant* and *thin glass* misperceived as *thing*. In the first case, the assimilated labial nasal /m/ aligned with the orthographic <m> in the misperceived words, wherein the subsequent /p/, which matched the intended target, may have prompted the extraction of the phonologically similar sequence '*impor*' and in turn

facilitated the selection of syntactically and semantically viable words (*important* or *unimportant*). In the latter case, the assimilation of /n/ to /ŋ/ was also reflected in the orthographic <g> in *thing*. However, the remaining segments of the phrase were disregarded in the erroneous responses, possibly indicating that once the phrase underwent nasal assimilation (i.e. resembling *thing glass*), no pragmatically viable alternative to the second word could be retrieved. These cases suggest that while *Regressive Assimilation* induces subtler phonological alterations compared to other CSPs, it can still pose challenges to L2 speech perception in certain contexts such as those illustrated above.

L1 Listener Comparison

As a comparison, L1 English listeners demonstrated ceiling-level performance, with a mean transcription accuracy of 96% in connected speech and 98% in citation form (cf. Figure 4.5). Among the few errors observed, most were attributable to typographical slips (e.g. *toy airplane* transcribed as **tou aeroplane*), word omissions (e.g. *be able* transcribed as *able*), or context-driven substitutions (e.g. *friends have* transcribed as *friends had* in '*I assume their friends have already left*'). These responses likely reflected momentary lapses in attention or reliance on global pragmatic cues, rather than perceptual difficulties induced by CSPs. Given that such errors were highly idiosyncratic and sporadic, they are not enumerated and discussed in further detail.

Nonetheless, a small number of cases did reveal the influence of CSPs, albeit with much lower frequency than among L2 listeners. Examples included *that's hers* misperceived as *that's us (Deletion of Onset), hid useful* misperceived as *hid juiceful* or *did juiceful* (*Palatalisation*), *thin glass* misperceived as *thing* or *thing glass* (*Regressive Assimilation*).

While these misperceptions were rare, they echoed patterns also observed in the L2 data, suggesting that certain CSP cases – particularly those that could result in phonetically similar and sometimes pragmatically viable real-word outcomes – may momentarily disrupt even L1 listeners' speech perception. Overall, however, the nature and frequency of errors among L1 listeners differed markedly from those of L2 listeners, with limited overlap in specific cases.

Summary

As illustrated above, these target CSPs appeared perceptually disruptive – particularly for L2 learners, who exhibited wide discrepancies in accuracy between connected speech and citation form. In contrast, L1 listeners maintained consistently high accuracy, with only rare and largely isolated misperceptions. Among L2 learners, the extent of difficulty varied across CSP categories: in some cases, (mis)perceptions were more directly driven by the phonology of the CSP itself, while in others, learners' responses appeared to be influenced by other factors such as pragmatics and grammar. The examples discussed above are representative of each CSP category, with a full list of individual stimuli and relevant sentential contexts provided in Appendix C. The post-hoc interpretations of the misperception patterns are informed by theoretical groundings, such as spoken word recognition (e.g. Marslen-Wilson & Welsh, 1978) and top-down and bottom-up processing (cf. Section 2.2.3 for a review of these approaches), critically underscoring the context-dependent and dynamic interplay between CSPs and L2 speech perception. I will now turn to consider the interventions and whether the training provided was able to facilitate L2 learners' perception of these CSPs.

6.2 RQ 2a & 2c Improvement in Accuracy-Based Intelligibility

Research Questions 2a and 2c were as follows:

RQ 2. What is the effectiveness of pedagogical interventions targeting CSPs on adult Mandarin-speaking learners' L2 speech perception?

RQ 2a. Do rule-based, noticing-based, and implicit approaches enhance the intelligibility (i.e. transcription accuracy) of connected speech for the L2 learners?

RQ 2c. If so, is one approach more effective relative to the others?

We conducted a rigorous intervention with a pre-test/post-test design with counterbalanced assessment stimuli and tightly controlled language input across intervention groups (cf. Sections 5.1.3 and 5.1.4). Consequently, potential biases related to stimulus order and quantity of targeted input during the interventions can be reasonably ruled out. The key finding was that participants demonstrated overall improvement in accuracy-based intelligibility measures when data from all three groups were aggregated. This main effect appeared to be driven by the experimental groups (the rule-based and noticing-based conditions), both of which showed increased mean scores from the pre-test to post-test, whereas the control group (the implicit condition) showed consistent mean scores over time (cf. Figure 5.23). However, despite numerically greater gains in the experimental groups than the control group, the differences in improvement between groups were statistically ambiguous. As such, no firm conclusions can be drawn regarding the relative effectiveness of the intervention conditions compared. Therefore, any discussion of group differences below should be regarded as tentative.

Since there was no evidence that differential improvement existed between groups, we do not have statistical evidence that the experimental conditions were effective compared

to a control; however, the overall pattern is at least consistent with the CSP-focused interventions being effective. To that extent, it aligns directionally with a number of prior studies (e.g. Kuo et al., 2016; Ahmadian & Matour, 2014; Ting & Kuo, 2012; Matsuzawa, 2006), though notably no prior studies (to my knowledge) have used the disparity score metric. Despite this, in our discussion we focus on the findings with this metric since (as noted in Section 5.2.2) it more precisely captures CSP-related difficulty.

In examination of group-specific effects, the rule-based group showed the numerically largest improvement in disparity scores. However, the evidence that the rule-based group's improvement exceeded that of the control (implicit) group, though trending in this direction, was in the ambiguous region (cf. Table 5.7 and Table 5.8). On the face of it, this finding appears to contrast with more definite claims in the literature. However, these claims are not always substantiated by the research designs or analyses used. For example, Ahmadian and Matour (2014) found a significant pre-to-post gain in an experimental group of Iranian EFL learners who received explicit CSP instruction, while a no-instruction control group showed no such improvement. On the basis of this, they claimed that the intervention was effective, but they did not test whether the gains of the groups were different. Thus, their conclusion rested on an over-interpretation of a non-significant *p*-value. In addition, in Ting and Kuo (2012), Taiwanese sophomore English majors who received explicit CSP instruction and practised via blank-filling exercises along with pop song lyrics significantly outperformed a no-instruction control group. Yet, notably these conclusions were drawn from comparisons of post-test group means without taking into account participants' ability prior to the intervention.

Some studies (e.g. Kuo et al., 2016; Rahimi & Chalak, 2017) incorporated pre-test

performance as a covariate using ANCOVA, an approach that adjusts for baseline differences and enhances the comparability of post-test outcomes. For instance, Kuo et al. reported that a group of Taiwanese eighth graders receiving explicit instruction improved more than a no-instruction control group, whereas a group engaging with structured communicative (interactive) oral tasks showed no advantage over the same control group. In an Iranian EFL context, Rahimi and Chalak similarly found a significant difference between an explicitly taught experimental group and a control group receiving no instruction. While ANCOVA adjusts post-test scores based on pre-test performance, it compares adjusted post-test outcomes rather than learning gains per se. In contrast, the mixed effects regression models employed in the current study focus directly on change over time by modelling interactions between group contrasts and assessments, while also accounting for individual variability in baseline differences (via random intercepts) and in learning trajectories (via random slopes) (cf. Section 5.2.2: *Bayesian Logistic Mixed Effects Regression Model*). These two approaches thus reflect distinct ways of treating baseline performance and serve conceptually different inferential purposes. Whereas ANCOVA evaluates adjusted post-test scores, the current models examine how much change occurs across assessments. This structure supports a more direct analysis of learning gains, particularly in samples that may be broadly classified as 'homogeneous' in initial proficiency.

Furthermore, several studies lacked a control group (e.g. Carreira, 2008; Matsuzawa, 2006), and those that did include one (e.g. Kuo et al., 2016; Ahmadian & Matour, 2014; Ting & Kuo, 2012; Rahimi & Chalak, 2017) typically used groups that followed their regular school curricula and received no CSP input or instruction. As such, their design did not allow them to tease apart whether it is implicit exposure or explicit instruction

that makes the difference, since the control group may not encounter any relevant input or the experimental group may get both.

By contrast, our study tactically exposed the control group to the same targeted CSP samples as those practised by the experimental groups in an implicit manner. Despite deliberate efforts to integrate CSP samples into the materials for the control group and to encourage their engagement with these samples through comprehension questions, the control group did not show meaningful improvement in disparity scores. Speculatively, two potential factors may account for this outcome. First, the intensive short-term nature of the intervention may have placed the implicit learning condition at a disadvantage (Ellis, 2009) – i.e. a longer duration of pedagogical treatment may have been necessary for L2 listeners to internalise the implicitly embedded CSP input and apply this knowledge to novel listening contexts. Second, given that the input for the control group was embedded within tasks that closely resembled conventional listening comprehension exercises (cf. Section 5.1.4: *Instructional Design*) – a learning method highly familiar to Taiwanese learners of L2 English – participants may have defaulted to extracting only the information necessary to answer comprehension questions, relying predominantly on more global cues (i.e. pragmatic relevance and plausibility), rather than engaging with the phonology of the CSP samples as effectively as the other two groups.

On the other hand, the absence of evidence of differences between the two experimental conditions in our study algins with Kuo et al. (2016). While their study compared explicit and communicative instruction and ours contrasted rule-based and noticing-based approaches, neither study found a clear advantage for one approach over the other. Similarly, in examining the effects of inductive instruction (guiding learners to compare

samples with and without CSPs to discern underlying patterns) versus deductive instruction (explicitly presenting CSP rules followed by drill-based practice), Kul (2016) reported no significant differences between the two groups in their dictation task performance; although they noted higher performance for the inductive group in an immediate post-test, the advantage did not persist over time. Overall, research comparing various pedagogical approaches – including rule-based vs. noticing-based (this study), explicit vs. communicative (Kuo et al., ibid.), inductive vs. deductive (Kul, ibid.) – have yet to reach any consensus on their relative effectiveness. One challenge lies in the variability across studies not only in the specific CSP categories targeted but also in the design of interventions (how CSPs are taught and what materials are utilised) and the degree of input control. Given these discrepancies, more rigorously controlled research is warranted to continuously refine our understanding of the effectiveness of CSP instruction.

A Tentative Classification of CSPs Based on Intervention Outcomes

Although the current study did not yield clear evidence of group differences for individual CSPs, numerical patterns suggest a potential three-tier classification of the investigated categories. From a phonological perspective, two of the six targeted CSPs – *Deletion of Onset* and *Palatalisation* – involve substantial structural alterations. *Deletion of Onset* entails the complete loss of a sound and *Palatalisation* merges two distinct sounds into a third intermediary one. The data indicated that the rule-based group exhibited the most numerically pronounced improvement for these two categories (b^{38} = 0.59 for *Deletion of Onset*; b = 0.58 for *Palatalisation*), whereas the other two groups showed near-zero

[38] All reported beta (b) coefficients in this and the following discussion sections refer to the effect of Assessment.ct in the Bayesian regression models for the relevant conditions, unless specified otherwise.

effects (cf. Figure 5.25). These patterns are consistent with an account in which CSPs involving more drastic phonological alterations may benefit particularly from explicit rule explanations, as L2 learners may otherwise struggle to infer and generalise these patterns to novel speech contexts by themselves.

On the other hand, *Reduction of Stop* and *CC-V Linking* involve subtler phonological alternations, as they affect only part of a target segment. *Reduction of Stop* results in the non-release of a word-final stop while maintaining its stop (air blocking) property, and *CC-V Linking* changes the syllabification of one segment while preserving all phonetic properties. Notably, *Reduction of Stop* was the only category where the control group (implicit condition) exhibited a trend towards improvement (b = 0.46), closely approximating the rule-based group (b = 0.49). For *CC-V Linking*, the noticing-based group demonstrated a stronger effect (b = 0.55) with the CIs *not* crossing zero, compared to the rule-based group (b = 0.34) with the CIs crossing zero. These trends potentially suggest that CSPs involving less drastic phonological alterations may be amenable to less explicit approaches, as learners may be able to discern and generalise such patterns through implicit exposure (as in *Reduction of Stop*) or structured guidance (as in *CC-V Linking*) without the need for metalinguistic explanations.

The final two categories – *Glide Insertion* and *Regressive Assimilation* – exhibited less obvious patterns. *Glide Insertion* introduces a new sound between two vowels, and while the rule-based (b = 0.19) and noticing-based (b = 0.17) groups showed comparable effects, both estimates were relatively small with great ambiguity. *Regressive Assimilation* entails the assimilation of a word-final alveolar nasal to the place of articulation of the following segment while retaining its nasal manner. Across all groups, the noticing-based group was

the only group that improved, but even for this group the effect (b = 0.12) was weak and highly ambiguous. Given the limited magnitude of these effects and their statistical uncertainty, the present data do not support any pedagogical approach for these two CSP categories. Studies with larger sample sizes are needed to determine whether more consistent patterns emerge.

It is important to emphasise that this proposed classification is based on post-hoc qualitative interpretations of data trends. The primary analytical results did not provide direct evidence for group differences in individual CSPs, and the observed patterns remain statistically ambiguous (cf. Section 5.2.3). Nevertheless, this tentative classificatory scheme, based on phonological mechanisms, provides a hypothesis that could be directly tested in future research. Note that most prior studies have not reported inferential statistics for individual CSPs but instead examined a number of CSPs collectively (e.g. Rahimi & Chalak, 2017; Ahmadian & Matour, 2014; Kul, 2016; Matsuzawa, 2006; Carreira, 2008). Among the few studies that have included effects for individual CSPs, Ting and Kuo (2012) provided partial support for the proposed classification, reporting that explicit instruction was effective for *Palatalisation* but not for *Deletion of Onset*. Conversely, Kuo et al. (2016) did not find explicit instruction to be effective for *Palatalisation*. Abe (2010) showed that explicit instruction combined with guided peer interaction (comparing CSP-present and CSP-absent samples) was more effective for *Palatalisation* than explicit instruction alone. However, isolated effects for the other CSP categories targeted in the current study have not, to my knowledge, been previously documented, thereby precluding meaningful comparisons. Given the scarcity of empirical evidence on the relative effectiveness of pedagogical approaches for individual CSP categories, larger-scale studies are needed to substantiate and refine this

proposed classification.

6.3 RQ 2b & 2c: Inverse Directional Effects in Comprehensibility Ratings

Research Questions 2b and 2c were as follows:

RQ 2. What is the effectiveness of pedagogical interventions targeting CSPs on adult Mandarin-speaking learners' L2 speech perception?

RQ 2b. Do rule-based, noticing-based, and implicit approaches enhance the comprehensibility (perceived ease of understanding) of connected speech for the L2 learners?

RQ 2c. If so, is one approach more effective relative to the others?

An unexpected finding of the current study was the inverse directional trends observed between intelligibility and comprehensibility measures. While overall participants demonstrated improvement in processing connected speech – as evidenced by increased transcription accuracy and disparity scores (the latter was reverse-coded such that increases reflect alleviation of CSP-induced difficulty) – they conversely rated connected speech stimuli as more difficult to understand following the interventions. Although the intelligibility and comprehensibility measures were correlated with each other at both pre-test and post-test (cf. Figure 5.30), generally in line with Munro and Derwing's seminal studies (1995; 1997), a divergence in the trajectories of these two measures has not been identified previously in the CSP literature, and as such, may tentatively be seen as a novel finding in this area.

To explore this further, if we focus on the rule-based group, which notably showed the clearest pattern of decline (b = -0.37, CIs[-0.68, -0.04]), one possible explanation is that

explicit metalinguistic training heightened these participants' awareness of phonological alterations, prompting them to engage in more conscious analysis when encountering similar CSPs in novel contexts (i.e. assessment stimuli). This increased self-monitoring may have raised their sensitivity to non-canonical (i.e. connected speech) pronunciations, potentially slowing down or interrupting their real-time speech processing. Consequently, participants may have perceived their understanding of stimuli as more 'deficient' and thus reduced their comprehensibility ratings. Meanwhile, the cognitive load imposed by unaccustomed phonological analysis may have diverted L2 listeners' attentional resources away from global pragmatic information, further compromising their perceived ease of understanding the overall meaning of the sentential stimuli.

However, this explanation alone does not explain the decline in the control group, which was only implicitly exposed to targeted CSP samples. When examined in isolation, this group also showed a numerical decline (b = -0.35, CIs[-0.80, 0.11]), though statistically uncertain, and critically there was no evidence that their outcomes differed from the experimental groups. This suggests that factors beyond metacognitive awareness of CSPs may have influenced these participants' judgement. One possibility relates to task design and cognitive fatigue. Unlike the experimental groups, who engaged in receptive listening training, the control group was required to answer open-ended comprehension questions during the intervention. This additional productive component may have imposed greater cognitive demands over time, resulting in attentional fatigue by the post-test, and thereby affecting participants' perceived ease of understanding and lowering their ratings.

Another potential factor concerns the use of British-accented intervention materials. Given that English instruction in Taiwan is predominantly based on American English –

a pattern reflected in participants' self-reported familiarity with L2 variety (95% reported the highest familiarity with American English). While the initial intent was to leverage the relative novelty of British-accented materials as a motivational tool for learning, extended exposure to such unfamiliar 'accented' speech may have inadvertently increased their awareness of accent-related phonetic/phonological features, thereby disrupting their processing fluency and reducing their perceived ease of understanding at the post-test. Some previous studies suggest that comprehensibility ratings of L2 speech are generally comparable between L1 listeners with or without experience of L2-accented speech (e.g. Kennedy & Trofimovich, 2008), while others report benefits of a shared language background between L2 listeners and L2 speakers (e.g. Foote & Trofimovich, 2018; Ludwig & Mora, 2017; Mora, 2022); nevertheless, the extent to which L2 listeners' familiarity with accents spoken by L1 speakers may influence their comprehensibility ratings has remained underexplored. Therefore, there seems to be an opportunity for future studies to explore whether CSP intervention effects on comprehensibility actually vary depending on L2 listeners' familiarity with different L1 accents (e.g. familiar American accent vs. unfamiliar British accent for Taiwanese learners). Additionally, some recent studies suggest that listeners' attitudes towards speakers – for example, through mechanisms of social priming (Taylor Reid et al., 2019; 2022) – can influence their comprehensibility judgements. This social factor could likewise be incorporated in pedagogical programmes in future CSP intervention research examining the impact of accent familiarity on L2 listeners' evaluations of speech produced by L1 speakers.

Importantly, the possibility that accent familiarity also underlies the rule-based group's decline in comprehensibility ratings must be acknowledged. Although their exposure was shorter and restricted to isolated sentences (as opposed to the extended discourses

presented to the control group), both groups received input in the same accent. In order to disentangle these potential explanations – accent familiarity versus metalinguistic processing – a larger-scale study is needed, specifically one with sufficient power to detect whether extended exposure leads to greater decline than explicit rule-based phonological training.

The overall pattern of decline in comprehensibility ratings (in the rule-based and control groups) may also echo the widely documented U-shaped trajectory in language development. An example of such a developmental trajectory specific to L2 comprehensibility was reported in Trofimovich et al. (2020), where L2 interlocutors' comprehensibility ratings of each other's speech (within pairs) tended to decrease temporarily before improving throughout a series of interactive tasks. In the current study, the observed decline in L2 listeners' comprehensibility ratings after the CSP interventions may suggest that these learners had not yet reached a threshold of familiarity with L1 connected speech at which increased comprehensibility might be expected, despite their improvement in the accuracy-based intelligibility measures. This highlights the need for future inquiries into whether CSP-related development follows a non-linear U-shaped pattern, particularly with respect to L2 listeners' comprehensibility ratings of L1 connected speech. Moreover, since CSPs can simultaneously affect segmental, prosodic and fluency features, this phenomenon may contribute to the broader body of research on how these phonological correlates interact with speech comprehensibility (see Saito, 2021, for a meta-analytic synthesis).

In contrast to the rule-based and control groups, the noticing-based group exhibited no evidence of change in comprehensibility ratings (b = -0.02, CIs[-0.34, 0.30]), likely

because this intervention condition fell between the rule-based and control groups. Without explicit instruction in CSP rules, it was possible that these participants were less inclined to engage in systematic phonological analysis during the post-test. Meanwhile, unlike the control group, they were not exposed to extended discourses in an unfamiliar accent, allowing them to focus more consistently on discerning phonological alterations (CSPs) during both the intervention and the post-test. As a result, participants in the noticing-based group may have been less susceptible to the competing cognitive demands that appeared to influence the other two groups. Nevertheless, it is important to acknowledge that the interaction terms did not yield conclusive evidence for differential effects across groups – neither the null nor the alternative hypothesis was clearly supported – and thus, these interpretations remain tentative.

Critically, this overall decrease in comprehensibility ratings should not necessarily be viewed as a deterrent to CSP instruction. Rather, it highlights a crucial pedagogical consideration: teachers and learners should be aware that a decline in comprehensibility (i.e. perceived ease of understanding) does not inherently indicate a decrease in accuracy-based intelligibility. As demonstrated in this study, the opposite – namely, improved transcription accuracy – may in fact be the case. In practice, such developmental patterns can be reframed as opportunities, for example, to reflect on potential (meta)cognitive or affective influences, as discussed above, on learners' perceived ease of understanding connected speech produced by L1 speakers, particularly in relation to their ability to recognise CSP-affected words.

It is also important to note that this regressive pattern in comprehensibility may not necessarily apply to learners with different demographic profiles or to other pedagogical

approaches. One implication from the findings is that, if the primary goal is to enhance the intelligibility of connected speech for L2 listeners, then incorporating CSPs into L2 instruction may be a beneficial pedagogical focus – even if this is accompanied by a potentially temporary decline in perceived ease of understanding. Future research could explore whether and under what conditions L2 listeners ultimately perceive L1 connected speech as more comprehensible and how this perceived ease of understanding might relate to any increases in accuracy-based intelligibility. Longitudinal studies encompassing extended CSP-focused interventions, multiple assessment time points, or sustained L1-L2 interaction may provide further insights into a potential U-shaped developmental pattern in L2 listeners' evaluations of speech produced by L1 speakers as well as their implications for L2 pedagogy.

6.4 RQ 2d: Broader Effects on General Listening Comprehension

Research Question 2d was as follows:

RQ 2. What is the effectiveness of pedagogical interventions targeting CSPs on adult Mandarin-speaking learners' L2 speech perception?

RQ 2d. Do the interventions targeting CSPs have a broader effect on the L2 learners' general listening comprehension?

Regarding the broader impact of the interventions on general listening comprehension, while a marginal main effect was observed (b = 0.25; CIs[-0.01, 0.52]), none of the intervention groups demonstrated statistically robust improvement when examined individually, nor were there meaningful group-by-assessment interactions. This falls short of providing robust evidence that the interventions enhanced general listening comprehension. These findings align with previous studies reporting no evidence of

improvement (Brown & Hilferty, 1986; Carreira, 2008) but contradict others that reported positive effects (e.g. Rahimi & Chalak, 2017). Differences in analytical methods across studies – such as paired sample t-tests (Carreira, 2008), one-way ANCOVA (Rahimi & Chalak, 2017), repeated measures ANOVA (Brown & Hilferty, 1986) – may have contributed to the discrepancies in results. However, none of the above studies directly tested interactions between intervention conditions and assessment time points to determine differential gains between groups (although the ANCOVA used by Rahimi and Chalak adjusted post-test scores based on pre-test performance, as discussed earlier in Section 6.2). Given that no evidence of improvement was identified in the current study (which employed Bayesian mixed effects regression modelling), the findings remain inconclusive regarding whether, or to what extent, CSP instruction may facilitate general listening comprehension.

Notably, although not statistically robust, the control group's CIs (b = 0.32, CIs[-0.04, 0.67]) were closer to the threshold than those of the experimental groups. This numerical gain for the control group is particularly noteworthy given the absence of evidence of improvement in any principal measure for this group. One tentative explanation is that the format of the listening comprehension practice may have aligned more closely with Taiwanese learners' prior experiences of listening instruction, possibly affording this group a relative learning advantage. Nonetheless, due to the lack of statistical support, such interpretations remain speculative and warrant further investigation with larger samples to detect potential group-specific effects and group-by-assessment interactions.

6.5 RQ 2e: Interrelationships among Intelligibility, Comprehensibility, and General Listening Comprehension

Research Question 2e was as follows:

RQ 2. What is the effectiveness of pedagogical interventions targeting CSPs on adult Mandarin-speaking learners' L2 speech perception?

RQ 2e. How do intelligibility, comprehensibility, and general listening comprehension measures correlate with each other before and after the interventions?

The correlation analyses indicate that intelligibility (transcription accuracy in connected speech, citation form, and the derived disparity scores), comprehensibility (perceived ease of understanding) and general listening comprehension (GLC) are positively correlated with one another (cf. Figure 5.30), and that these correlations remain stable before and after the interventions (cf. Table 5.15). This represents a novel contribution to the existing body of research, particularly as prior CSP intervention studies have rarely examined all three variables simultaneously. Most studies have focused on intelligibility, and some have included general listening comprehension, but none have concurrently incorporated comprehensibility into their analyses alongside the other two variables.

The positive relationship found between connected speech transcription accuracy and general listening comprehension corroborates Wong et al.'s (2017) model (cf. Figure 2.3 in Literature Review), which identified a direct regression path from intelligibility (labelled as 'reduced forms perception' in their study) to listening comprehension. However, the inclusion of comprehensibility in the current study potentially introduces an additional influential variable not accounted for in Wong et al.'s model. While further investigation is needed to determine the exact magnitude of its direct or mediating effects, comprehensibility emerges as a relevant factor, given its significant correlations with both

intelligibility and GLC.

More broadly, the interrelationships among these three variables reflect the dynamic interactions between cognitive mechanisms, as outlined in Vandegrift and Goh's (2012) combined model (cf. Figure 2.1 in Literature Review), which builds upon Levelt (1993) and Anderson (1995). According to this model, intelligibility – specifically, the transcription of CSP-affected words – primarily pertains to the 'perception' stage and partially the 'parsing' stage. The former involves the formation of phonetic representations from acoustic-phonetic signals, and the latter incorporates phonological decoding and lexical selection (cf. Section 6.1 for the discussion of linguistic details of CSPs and relevant examples). In contrast, comprehensibility engages a higher-order metacognitive dimension, as listeners reflect on and evaluate their ease of understanding. Since participants provide comprehensibility ratings immediately after transcribing each stimulus, these ratings inherently encapsulate broader speech-processing experience at the metacognitive level, on top of 'perception' and 'parsing'.

If pedagogical interventions facilitate 'perception' and 'parsing' (i.e. improvement in intelligibility), enhancement in subsequent comprehension – termed the 'utilisation' stage in the model – would be expected. However, while overall gains in intelligibility and its positive correlations with the GLC may partially support this assumption, the group-specific and interaction effects remain ambiguous. Notably, conditions which demonstrate clearer numerical gains in intelligibility do *not* necessarily correspond to a larger improvement in listening comprehension (cf. Table 5.7 and Table 5.13). Furthermore, as discussed above in Section 6.3, post-intervention comprehensibility ratings even reveal a regressive effect despite improvement in intelligibility. These

findings suggest that although intelligibility, comprehensibility, and GLC are significantly interrelated, their interactions across different stages of (L2) speech processing are complex. A critical pedagogical question that arises is why an increased ability to recognise CSP-affected words does not directly translate into higher comprehensibility ratings or enhanced listening comprehension. This likely reflects the influence of factors such as heightened phonological awareness, U-shaped developmental trajectories, or alignment with familiar versus unfamiliar learning formats, as previously discussed (cf. Sections 6.3 and 6.4), or possibly the presence of other contributors yet to be identified. Longitudinal studies administering a wider range of measures are needed to further explore the role of CSPs within this framework and refine our understanding of their effects on and interactions with different stages of L2 speech processing.

6.6 Practical Considerations for Teaching CSPs

Having discussed the interpretations of the study's findings in response to the research questions (Sections 6.1 – 6.5), this section outlines four key pedagogical considerations for practitioners. These practical implications are intended to connect theory to practice and support teachers in integrating CSPs into L2 instruction. Specifically, they address (1) the importance of practitioners' metalinguistic awareness; (2) considerations of language variety and accent in listening input; (3) strategies for curating authentic materials; and (4) the applicability of CSP instruction to diverse learner populations and contexts.

6.6.1 Practitioners' Metalinguistic Awareness

The diagnostic data from Phase 1 clearly indicate that different CSP categories exert varying degrees of influence on L2 speech perception. A pedagogical implication from these findings is that teaching practitioners should be aware of not only the impact of

CSPs on learners' speech perception but critically the differential impact of different CSP categories. This awareness is crucial for deciding which categories may warrant prioritisation for treatment, particularly as in teaching practice, resources such as time and instructional autonomy are often constrained, such that it is often impractical to address all CSP categories comprehensively within an instructional programme.

In this regard, a prerequisite for any potentially effective treatment may involve teachers developing a nuanced understanding of CSPs. In particular, it may be important for teachers to move beyond the commonly oversimplified notions of CSPs found in many existing teaching materials – such as general references to 'linking' or 'casual pronunciation' – and instead recognise the diverse phonological mechanisms that underpin connected speech. Moreover, the Phase 2 study tentatively suggests that explicit rule-based instruction may hold promise in supporting L2 learners' speech perception compared to solely implicit input, which – if corroborated in larger-scale studies – would point to the potential importance of teachers possessing strong metalinguistic knowledge of CSPs. Practically, even when using pedagogical approaches that foreground implicit or noticing-based instruction, as explored in this study, teachers may still need a well-developed understanding of CSPs (including their underlying phonological mechanisms and categorisation). This knowledge can support the curation of suitable sample materials and facilitate learners' systematic and meaningful exposure to the target features. When considered alongside prior research, the overall findings of this study imply that improvement in L2 learners' processing of connected speech may be more effectively fostered through targeted CSP input than through incidental and unsystematic exposure. Consequently, in L2 instruction, it may be beneficial for teachers to be equipped with sufficient metalinguistic knowledge to design or adapt instructional materials that are

appropriately aligned with learners' developmental needs.

With respect to intelligibility in particular, the empirical evidence – reflected in the numerically greater gains observed in the experimental groups – suggests potential benefits of implementing targeted instruction in CSPs. These findings, however, are not intended to be a prescriptive mandate for replacing existing instructional methods. Instead, they provide an evidence-based foundation upon which teachers can make informed pedagogical decisions, adapting and integrating CSP instruction within their specific teaching contexts. The overarching goal is not the wholesale adoption of any single pedagogy but rather the incorporation of research-informed strategies that may complement and enhance existing curricula.

6.6.2 Language Variety and Accent in Listening Input

The implications of this study must also be considered in light of the English varieties and accents[39] used in assessment stimuli and intervention materials. The stimuli were produced in mid-southern UK English and the intervention materials featured General UK English (cf. Section 4.1.3 and Section 5.1.4). Both were less familiar to participants in the current study, 95% of whom identified American English as their most familiar

[39]It is acknowledged that the terms *variety* and *accent* are closely related. Operationally, in the current context, variety is used to refer to a broader set of linguistic features represented in speech stimuli and intervention materials, encompassing not only phonetic/phonological characteristics (i.e. accent) but also potentially lexical, syntactic, and pragmatic features. Accent, on the other hand, is used to denote phonetic/phonological variation associated with the variety spoken by individuals. While the former serves as a more general descriptive label, the latter is particularly relevant in discussing perceptual responses to speech materials – as the literature often isolates accentedness as a variable partially correlated with intelligibility and comprehensibility (e.g. Munro & Derwing, 1995), or frames it as an alternative, albeit contested, focus for L2 pronunciation instruction (e.g. Levis, 2005), among other possibilities.

English variety. Although the inherent variability of language varieties – a factor that is difficult to control due to high speaker idiosyncrasies, particularly in CSP research – means that similar learning outcomes may not be guaranteed when different English varieties are used, the current study highlights that even when L2 learners are exposed to a less familiar accent, targeted instruction may still promote gains in intelligibility when processing speech produced by L1 speakers.

This consideration of accent recalls studies by Munro and Derwing (1995, 1997), which demonstrated that L1 listeners' perception of the accentedness of L2 speech correlates with their transcription intelligibility scores and comprehensibility ratings. While a similar relationship may reasonably be expected in L2 listeners' perception of L1 speech, their evaluations of accentedness are likely to show lower reliability. In this regard, a greater number of studies documenting these related variables simultaneously would allow for comparisons of L2 listeners' transcription accuracy and comprehensibility ratings across different accents in the target languages.

At the same time, it is important to note that 'universality' was one key criterion in the CSP selection process in Phase 1 (cf. Section 4.1.2), ensuring that the CSP categories investigated in this study are not exclusive to a single English variety. Therefore, while it was necessary to select specific varieties for empirical investigation, the pedagogical implications of teaching CSPs may remain relevant across different English varieties. Nonetheless, to develop a more comprehensive understanding, further research is needed to examine the precise interplay between L2 listeners' familiarity with varieties used by L1 speakers and the effectiveness of CSP-focused interventions in enhancing speech intelligibility and comprehensibility.

6.6.3 Curation of Authentic Materials

Beyond understanding what categories of CSPs to teach and how to teach them, access to systematic and reliable teaching materials remains a prominent challenge. This issue was particularly evident during the design and preparation of the intervention study, as existing resources were insufficient. In response, this study demonstrates a practical approach to curating materials using YouGlish, an internet-mediated platform that allows users to search for specific words and phrases within authentic audio-visual content (cf. Section 5.1.4: *Instructional Design* and see sample materials in Appendix E). The approach offers two advantages: first, it provides a convenient way of sourcing CSP examples; second, it reinforces the argument for authenticity in pronunciation-related instruction. Given that CSPs occur naturally in spontaneous communication, authentic materials are not only beneficial but likely necessary for effective instruction.

Most content retrievable from YouGlish is not originally created for language teaching purposes, thereby reflecting naturalistic pronunciation patterns that learners are likely to encounter in real-world interactions. This contrasts with scripted or decontextualised samples often created specifically for language instruction, which may not fully capture the phonological features of naturally occurring speech. This study tentatively suggests that, when carefully selected and strategically embedded into instructional materials, CSP samples extracted from open and authentic sources may be beneficial for L2 listeners' speech perception even within the context of short-term intensive interventions. Crucially, this approach does not require a fully developed textbook or custom-made audio/video recordings produced by teachers from scratch, making it a practical and scalable solution for integrating CSP instruction into curricula. Notably, as YouGlish functions across a

range of languages, this approach also holds potential for adaptation in L2 learning contexts beyond English.

In addition, an important pedagogical consideration concerns the potential role of high-variability phonetic training (Iverson et al., 2005). YouGlish provides an accessible means of curating materials featuring diverse speakers and contexts, which could potentially enhance L2 speech perception by leveraging variability effects (e.g. Logan et al., 1991; Lively et al., 1993; 1994). However, while earlier research supports the facilitative role of variability, more recent studies suggest that its benefits may be conditional and less consistent (e.g. Giannakopoulou et al., 2017; Brekelmans et al., 2022; Mora et al., 2022; Mora-Plaza et al., 2022). Given the mixed evidence, the material curation approach utilised in this study points to a strand of future research into variability effects on the learning of CSPs – specifically, how exposure to or instruction using multi talkers and/or varied communicative contexts may influence the effectiveness of CSP-focused interventions, particularly in contrast to low-variability input.

6.6.4 Applicability to Diverse Learner Populations and Contexts

This study highlights the persistent difficulty L2 learners face in processing even familiar lexical items when affected by CSPs. A substantial discrepancy was evident between participants' high transcription accuracy for citation form items – further supported by strong lexical-semantic knowledge, indicating participants' familiarity with the target words (see Appendix F for the analysis of this supplementary measure) – and their significantly lower accuracy when the same words were produced in connected speech. These findings suggest that intelligibility is not solely a function of lexical knowledge and recognition of citation forms, but it is critically contingent on how words are realised

within connected speech. Crucially, this challenges the assumption that basic vocabulary items and phrases – those typically classified as within L2 leaners' general proficiency level – must be inherently easy to process in connected speech. Notably, even advanced participants in Phase 1 (with C1 or higher proficiency) struggled with beginner- and intermediate-level words when these words were produced in connected speech, further demonstrating the pronounced impact of CSPs and underscoring the potential value of pedagogical interventions that address CSPs for learners across proficiency levels.

In light of these results, targeted instruction in CSPs need not be restricted to advanced learners but could be systematically incorporated into L2 language curricula for learners across a broader range of proficiency levels. While this study targeted upper-intermediate to advanced learners, prior empirical research similarly indicates that beginner- to low-intermediate-level learners also experience significant difficulties with CSPs (e.g. Abe, 2010; Rahimi & Chalak, 2017), as might be expected given their often limited exposure to naturalistic spoken input. Since CSP-related challenges appear to persist throughout the L2 learning trajectory, integrating CSPs into L2 instruction from earlier stages is likely to be more pedagogically effective than expecting learners to decipher these patterns independently and delaying treatment until later in their L2 development.

Additionally, difficulties with CSPs do not appear to be limited to learners from a particular L1 background. Previous studies have documented such challenges among speakers of Mandarin (this study; Kuo et al. 2016; Ting & Kuo, 2012), Japanese (e.g. Abe, 2010; Matsuzawa, 2006), Persian (e.g. Ahmadian & Matour, 2014) and Polish (e.g. Kul, 2016). Collectively, these findings suggest that CSPs present widespread difficulties in L2 learning, which may point to the potential applicability of including CSPs within

pedagogical practice across diverse linguistic contexts. However, further research is needed to determine how specific pedagogical approaches can be more effectively adapted for different learner populations.

As with all pedagogical implications, teachers should exercise professional judgment, leveraging their metalinguistic knowledge to tailor instructional approaches to learners' specific needs and curricular objectives. Importantly, given the dynamic nature of CSPs in spontaneous connected speech, empirically tested intervention approaches (even if subsequent research demonstrates their efficacy) should not be viewed as a rigid principle but rather as adaptable pedagogical frameworks that can be flexibly integrated into diverse learning contexts.

Chapter 7: Conclusions

7.1 Summary of Findings and Claims

This study set out to understand the challenges that connected speech processes (CSPs) impose on L2 listeners, and to explore ways of addressing these challenges. It was motivated by the premise that different CSP categories may exert varied degrees of impact, and that effective pedagogical interventions should prioritise those presenting the greatest challenges. To this end, a stimulus inventory was developed systematically comprising 116 sets of items exemplifying ten categories of CSP. The diagnostic experiment (Phase 1) demonstrated that the ten CSP categories investigated significantly affected L2 participants' transcription accuracy and, as expected, did so to differing extents. Critically, by comparing the accuracy with which participants transcribed the same lexical items produced in citation form and connected speech, six particularly challenging CSP categories were identified as priorities for interventions: *Deletion of Onset*, *Glide Insertion*, *Reduction of Stop*, *CC-V Linking*, *Palatalisation*, and *Regressive Assimilation*.

Building on these findings, the intervention study (Phase 2) sought to explore effective approaches for teaching CSPs under highly controlled input conditions. Three distinct intervention programmes – a rule-based condition, a noticing-based condition, and an implicit condition (serving as a control group) – were developed, each incorporating identical CSP samples into the instructional materials. The effectiveness of these interventions was assessed across three outcome variables: intelligibility (transcription accuracy), comprehensibility (perceived ease of understanding), and general listening comprehension. Results indicated an overall improvement in intelligibility, yet while there were numerical gains in general listening comprehension, they were not statistically robust. Unexpectedly, an overall decline was observed in comprehensibility ratings. In

other words, although participants improved in recognising CSP-affected words following the interventions, they subjectively perceived connected speech as more difficult to understand. However, Bayesian regression models and Bayes factor analyses provided ambiguous statistical evidence regarding group-by-assessment interaction terms, leaving the relative advantage of the different pedagogical approaches inconclusive. Although numerical trends suggest a possibility of greater improvement for the rule-based condition in accuracy-based intelligibility, definitive conclusions could not be drawn.

Three principal claims arise from the two cumulative studies: (i) CSPs impact the perception of even familiar words by high-proficiency L2 listeners to a much greater extent than is the case for L1 listeners; (ii) not all CSP categories impinge on L2 speech perception to the same degree, underscoring the necessity of systematic pedagogical prioritisation; and (iii) tentatively, pedagogical interventions involving structured CSP instruction or exposure may be beneficial for L2 speech perception (in terms of intelligibility), though the relative advantage of different conditions remains inconclusive. As this doctoral research progressed, I developed an increasingly nuanced understanding of both its contributions and the avenues for future exploration. These reflections are elaborated in the subsequent sections.

7.2 Contributions of the Research

Research Design and Linguistic Measures

This research introduces a diagnostic phase that systematically evaluates the relative impact of different CSP categories, thereby empirically justifying the selection of CSPs targeted for interventions. This mitigates the risk of arbitrary selection and facilitates evidence-based pedagogical prioritisation. Additionally, by calculating the discrepancy

between citation form and connected speech transcription accuracy, this research offers a precise assessment of CSP-specific difficulty. This approach contrasts with prior studies which primarily examined connected speech without accounting for baseline citation form performance. Critically, the consistent analysis of individual participants' disparity scores minimises the possibility of overestimating the difficulty associated with CSPs, by controlling for instances in which listeners may have struggled with the citation form itself. Moreover, the inclusion of comprehensibility ratings reveals an unexpected trend of decline in learners' perceived ease of understanding, despite an overall improvement in accuracy-based intelligibility metrics. This highlights the dynamic, non-linear nature of the perceptual development in L2 learners and prepares teaching practitioners for a potential 'transitional' phase in learner progress.

Intervention Programmes

This research implemented rigorous control over language input across the intervention groups – an important factor often overlooked in prior CSP studies. By embedding identical targeted samples in each intervention programme, this design enabled more reliable comparisons between groups and reduced potential assessment biases. Specifically, unlike studies where only experimental groups receive CSP instruction while control groups receive no CSP exposure, the current intervention framework ensured that all groups engaged with the same linguistic items, albeit through distinct pedagogical approaches. Additionally, within each programme, the sequencing of instructional sessions was counterbalanced to minimise potential recency effects – another aspect rarely addressed in previous CSP intervention studies. Together, these considerations allowed for a more valid attribution of learning outcomes to the intervention conditions rather than to discrepancies in targeted input, thereby supporting more reliable

evaluations of instructional effectiveness.

Statistical Analysis

The Bayesian mixed effects regression models employed in this research account for both individual baseline differences (via random intercepts) and individual developmental trajectories (via random slopes). The incorporation of Bayes factors as an additional inferential statistic further enhances the rigour of the analyses by distinguishing between evidence supporting the null hypothesis and evidence that is inconclusive. This distinction is particularly informative in the current research context for interpreting key interaction terms – those that could have indicated a relative advantage of one condition over another, but did not reach conventional statistical thresholds. Notably, such distinction is not achievable with the frequentist approach and remains largely unexplored in prior CSP intervention studies. Furthermore, in contrast to earlier research in which interaction effects (i.e. direct comparisons of differential gains between groups) have rarely been reported, this study presents a more comprehensive examination of main, group-specific, and interaction effects. The analytical depth afforded by these techniques contributes to a more nuanced understanding of the effectiveness of different intervention conditions.

Practical Applications

On a practical level, this research demonstrates how authentic spoken materials can be curated via a freely accessible internet-mediated platform and integrated into diverse pedagogical approaches. Specifically, by embedding these materials within rule-based, noticing-based, and implicit conditions, it offers a replicable protocol for adapting real-world language input to suit different instructional methods. From the learner's perspective, engagement with realistic speech helps narrow the gap between classroom

language and natural spoken interaction, particularly in contexts where access to authentic and well-developed CSP instructional resources is limited. For teachers, the proposed methods are highly scalable and adaptable to diverse educational settings, including those shaped by institutional constraints or standardised curricula. Crucially, this study not only presents a conceptual framework for material development but also provides concrete, ready-to-adapt resources that exemplify how CSP instructional programmes can be practically implemented.

7.3 Future Directions

Although this two-phase research identified the varied impact of different CSP categories on L2 speech perception and evaluated the effectiveness of distinct pedagogical interventions, some questions remain unanswered. This section synthesises the emerging issues into four key areas, positioning them within the broader context of L2 speech perception research and identifying them as opportunities for future inquiry.

Bayes Factors and Statistical Power

The use of Bayes factors is advantageous for distinguishing evidence for the null hypothesis (H_0) and inconclusive evidence, thereby mitigating the limitations inherent in the frequentist null hypothesis significance testing. However, the validity of this approach is contingent on the specification of priors (cf. Section 5.2.2: *Bayes Factor Analysis*), raising concerns about potential biases. In the absence of pre-registered piloting data, this research employed the motivated-maximum approach (Silvey et al. 2024). Looking ahead, the current empirical study provides the data needed to compute informative priors which can be used in future larger-scale intervention studies. These empirically derived priors will facilitate more robust Bayes factor comparisons, particularly for evaluating

interaction parameters that may reflect differential effects across intervention conditions. Moreover, as one of the first CSP intervention studies to employ Bayes factors, the analytical framework established in this work may serve as a reference point for subsequent studies adopting the Bayesian approach.

In addition, while the principal measures indicated overall main effects of assessment time across conditions, there was no evidence for the interactions which would suggest that the extent of improvement (as in intelligibility) or decline (as in comprehensibility) depended on conditions – though notably, the Bayes factors fell within the ambiguous region, supporting neither the presence of interaction effects (H_1) nor the absence thereof (H_0). Although power analysis was conducted in advance, unforeseen factors – such as the exclusion of specific CSP categories and stimulus items – reduced the available data for inferential modelling. These exclusions adhered to the methodological criteria to minimise selection bias, but inevitably diminished statistical power, particularly for the subset analysis of individual CSPs. To address this limitation, I intend to build on the protocols established in this research and use simulation to estimate the statistical power required to detect both interaction effects and differential effects across individual CSP categories. This will support the refinement and validation of the proposed CSP classificatory scheme (cf. Section 6.2), which is tentatively grounded in numerical patterns of group-specific improvement (cf. Figure 5.25).

Stimuli and Intervention Materials

This stimulus inventory developed in this research (116 experimental items and 58 distractors) offers a foundation for further expansion following the same piloting procedures. Future longitudinal investigations could involve extended intervention

periods and incorporate multiple post-test sessions to examine the retention of any intervention effects over time. Furthermore, given the overall directional effects observed under highly controlled conditions, it would be valuable to adapt the intervention materials for use in classroom-based studies to examine whether comparable effects emerge. There is also an opportunity to explore alternative CSP pedagogical approaches, such as integrating educational technology or combining instructional techniques examined in this and prior studies, potentially advancing our knowledge of L2 speech perception pedagogy.

Additionally, since CSPs in this study were permitted to occur at flexible positions within sentential stimuli (excluding sentence-initial and sentence-final positions) in order to enhance authenticity, future studies could investigate whether the syntactic positioning of target CSP items influences L2 learners' perception of connected speech. Other controlled manipulations of training input – such as high-variability (e.g. diverse contexts and multiple speakers) versus low-variability (e.g. fixed contexts and single speaker) – may likewise yield insights into optimal training conditions. Lastly, studies incorporating a broader range of L2 varieties – beyond the UK English varieties featured in this research – could further elucidate role of accent familiarity in CSP intervention effects.

Assessment Framework

The overall declining pattern in comprehensibility ratings following the interventions suggest a potential affective dimension that may warrant strategic considerations by teaching practitioners when addressing CSPs. In light of this, it may be worthwhile to explore whether the effectiveness of CSP interventions in improving comprehensibility for L2 listeners is associated with other self-efficacy-related measures, such as L2 grit

(e.g. Teimouri et al., 2020), anxiety (e.g. Saito et al., 2018) and motivation (e.g. Nagle, 2018). Moreover, the interrelationships between intelligibility, comprehensibility, and general listening comprehension observed in this study, could inform broader language assessment frameworks and speech processing models.

Beyond the quantitative measures employed, future studies integrating qualitative methods – such as post-hoc recall interviews and think-aloud tasks – could provide deeper insights into the cognitive mechanisms underlying L2 speech perception. Such methods may be particularly valuable for elucidating the reasons behind erroneous responses that do not appear readily attributable to the influence of CSPs (cf. discussion on misperception cases in Section 6.1). Additionally, with the increasing prominence of L2 speaker-L2 listener interactions, the present assessment framework could be extended to examine interactions among L2 users with varying characteristics – such as proficiency levels, cultural-linguistic backgrounds, educational stages – and potentially across different target languages.

Pedagogical Focus of Listening Instruction

A broader implication emerging from this research concerns the shortcomings of relying predominantly on lexical coverage when selecting spoken materials for L2 listening instruction. The findings presented herein demonstrate that even high-frequency, familiar lexical items – generally presumed to be readily accessible to learners – can become unintelligible when affected by CSPs, thereby disrupting lexical recognition (cf. Section 3.3.2: *L1 Lexical-Semantic Knowledge Test* (LSK) for lexical-level control; Section 6.6.4 for the discussion of learner population and connected speech intelligibility). This calls into question the prevailing pedagogical assumption that high lexical coverage alone

constitutes a sufficient basis for processing L2 speech.

To take the context of English secondary schools as an illustrative example, recent curriculum guidance from Ofsted (2021) reiterates the centrality of high-frequency vocabulary, phonics, and grammar as foundations of language learning. The report highlights the importance of 'automatic and fluent recall' (p. 16) of vocabulary across modalities of language use and posits that 'with time and practice, knowledge of phonics, grammar and vocabulary becomes automatised … With this, learners can understand longer written texts and spoken discourse' (p. 20). While these objectives broadly align with general principles of language acquisition, the guidance does not appear to address the phonological variability inherent in spontaneous connected speech. Notably, even in its consideration of the 'real-time nature of speech (e.g. its speed and transitory nature)' (p. 26) as a factor in the selection of authentic spoken texts, the document does not explicitly identify CSPs as a distinct source of processing difficulty, particularly for learners who might otherwise be lexically equipped to understand a given spoken text.

Although the current study does not specifically examine text selection, its methodologies and findings nonetheless contribute a critical nuance to this pedagogical landscape. They suggest that lexical coverage, while necessary, may not fully and reliably predict speech intelligibility for L2 listeners in the absence of sensitivity to CSPs. In practical terms, this underscores the need for more refined criteria in the selection and design of materials for listening instruction – criteria that extend beyond lexical profiling to systematically account for phonological variability in spoken input. Future research could explore how instructional materials and curricular frameworks might more precisely reflect the demands of real-world speech – for instance, by comparing training conditions with and

without authentic input, varying the degrees of explicitness or implicitness in CSP instruction, and examining the benefits of supplementary supports such as transcripts or textual enhancement. Such inquiry may potentially inform the development of more comprehensive curriculum designs and deepen our understanding of L2 speech perception, thereby enabling more targeted and effective pedagogical interventions.

7.4 Final Remarks

The impetus for this doctoral research emerged during my TESOL training. At the time, my primary aspiration was to become a teaching practitioner, not a researcher. However, my fascination with speech science soon took hold, as I became increasingly humbled by the cognitive and perceptual intricacies involved in processing spoken languages – particularly a second language. More importantly, I was struck by the extent to which real-life pronunciations diverge from the 'canonical' forms we are expected to teach.

These dynamic alterations in connected speech raise the question of whether it is realistic to assume that L2 users can acquire such a system (or any other) as 'naturally' and 'smoothly' as L1 users. Motivated by my experiences as both a learner and a teacher, I sought to approach this issue systematically – first through diagnosing the problem, then exploring potential pedagogical solutions beyond what the existing literature had already offered, in the hope of advancing our understanding. Throughout this doctoral project, from the development of stimuli and intervention programmes to experimental design and statistical modelling, I have gained not only technical knowledge and skills but, more critically, developed a deeper appreciation for both the possibilities and limitations inherent in every stage of research.

Coming from a teaching background and now moving into academia, I have become acutely aware of the challenges in bridging the gap between these two supposedly connected educational spaces. As teaching practitioners, it may be instinctive to seek immediate and convenient solutions to pedagogical issues – an inclination I once shared. Over time, though, I have come to recognise how rarely such straightforward answers exist, if at all. In investigating CSPs – an often oversimplified yet critically important aspect of speech perception – this doctoral research does not mark an endpoint for me but rather a step in a journey of broader, ongoing inquiry. While the field has made significant strides, fully demystifying CSPs requires a sustained commitment to questioning, refining, and expanding upon existing scholarship. It is hoped that with collaborative efforts from academic and educational communities, incremental advancements – in theoretical conceptualisations, empirical findings, and pedagogical solutions – will continue to shape the field.

References

Abe, H. (2010). Form-focused Instruction in L2 Pronunciation Pedagogy: The effect of Negotiation of Form in a Japanese classroom. In K. Dziubalska Kołaczyk, M. Wrembel, & M. Kul (Eds.), *New Sounds 2010: Proceedings of the 6th international symposium on the acquisition of second language speech* (pp. 1–6).

Ahmadian, M., & Matour, R. (2014). The effect of explicit instruction of connected speech features on Iranian EFL learners' listening comprehension skill. *International Journal of Applied Linguistics & English Literature*, *3*(2), 227–236.

Alameen, G., & Levis, J. M. (2015). Connected Speech. In M. Reed & J. M. Levis (Eds.), *The handbook of English pronunciation* (pp. 157–174). Wiley.

Anderson, J. R. (1995). *Cognitive psychology and its implications* (4th ed.). Freeman.

Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning, 42*(4), 529–555. https://doi.org/10.1111/j.1467-1770.1992.tb01043.x

Anderson-Hsieh, J., Riney, T., & Koehler, K. (1994). Connected speech modifications in the English of Japanese ESL learners. *IDEAL*, *7*, 31–55.

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods, 52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Avery, P., & Rice, K. (1989). Segment structure and coronal underspecification. *Phonology, 6*, 179–200. https://doi.org/10.1017/S0952675700001007

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). *Fitting linear mixed-effects models using lme4*. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Theoretical and methodological issues* (pp. 171 – 203). York Press.

Boersma, P., & Weenink, D. (2024). *Praat: Doing phonetics by computer* (Version 6.4.12) [Computer software]. University of Amsterdam. https://www.fon.hum.uva.nl/praat/

Bowen, J. D. (1975). *Patterns of English pronunciation*. Newbury House.

Bowen, J. D. (1976). Current research on an integrative test of English grammar. *RELC Journal*, *7*(2), 30–37.

Brekelmans, G., Lavan, N., Saito, H., Clayards, M., & Wonnacott, E. (2022). Does high variability training improve the learning of non-native phoneme contrasts over low variability training? A replication. *Journal of Memory and Language*, *126*. https://doi.org/10.1016/j.jml.2022.104352

British Educational Research Association. (2018). *Ethical guidelines for educational research* (4th ed.). British Educational Research Association. https://www.bera.ac.uk/researchers-resources/publications/ethical-guidelines-for-educational-research-2018

Brown, C. A. (1998). The role of the L1 grammar in the L2 acquisition of segmental structure. *Second Language Research*, *14*(2), 136–193.

Brown, G. (2013). *Listening to spoken English* (2nd ed.). Routledge.

Brown, J. D., & Hilferty, A. (1986). The effectiveness of teaching reduced forms of listening comprehension. *RELC Journal*, *17*(2), 59–70.

Buck, G. (2001). An overview of listening comprehension. In *Assessing listening* (pp. 1–30). Cambridge University Press.

Burley-Allen, M. (1995). *Listening: The forgotten skill* (2nd ed., Wiley self-teaching

guides). Wiley.

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28. doi.org/10.18637/jss.v080.i01

Cahill, R. (2006). Teaching reduced interrogative forms to low-level EFL students in Japan. In J. D. Brown & K. Kondo-Brown (Eds.), *Perspectives on teaching connected speech to second language speakers* (pp. 99–125). National Foreign Language Resource Center, University of Hawai'i at Manoa.

Cambridge University Press & Assessment (2025). *English vocabulary profile online.* English Profile: The CEFR for English. Retrieved June 10, 2025, from https://englishprofile.org/?menu=evp-online

Cambridge University Press & Assessment. (2025). *IELTS (International English Language Testing System)*. Cambridge English. https://www.cambridgeenglish.org/exams-and-tests/ielts/

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1–47. https://doi.org/10.1093/applin/I.1.1

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 2–14). Longman.

Carpenter, B., Gelman A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, 76*(1). doi.org/10.18637/jss.v076.i01

Carreira, J. M. (2008). Effects of teaching reduced forms in a university preparatory course. In K. Bradford Watts, T. Muller, & M. Swanson (Eds.), *JALT 2007: Conference proceedings* (pp. 200–207). JALT.

Celce-Murcia, M., Brinton, D., & Goodwin, J. (2010). *Teaching pronunciation : A course book and reference guide* (2nd ed.). Cambridge University Press.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2002). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Taylor & Francis Group.

Council of Europe. (2025). *Common European Framework of Reference for Languages (CEFR)*. https://www.coe.int/en/web/common-european-framework-reference-languages

Cox, N. J. (2008). Speaking Stata: correlation with confidence, or Fisher's z revisited. *The Stata Journal*, *8*(3), 413–439. https://doi.org/10.1177/1536867X0800800307

Crawford, M., & Ueyama, Y. (2011). Coverage and instruction of reduced forms in EFL course books. *The Language Teacher*, *35*(4), 55–61.

Creative Commons. (n.d.). *Creative Commons*. https://creativecommons.org/

Cross, J., & Vandergrift, L. (2015). Guidelines for designing and conducting L2 listening studies. *ELT Journal, 69*(1), 86–89.

Crowther, D., Trofimovich, P., & Isaacs, T. (2016). Linguistic dimensions of second language accent and comprehensibility: Nonnative listeners' perspectives. *Journal of Second Language Pronunciation, 2(*2), 160–182. https://doi.org/10.1075/jslp.2.2.02cro

Crowther, D., Trofimovich, P., Saito, K., & Isaacs, T. (2018). Linguistic dimensions of l2 accentedness and comprehensibility vary across speaking tasks. *Studies in Second Language Acquisition*, *40*(2), 443–457. https://doi.org/10.1017/S027226311700016X

Cruttenden, A. (2014). *Gimson's pronunciation of English* (8th ed.). Routledge.

Cutler, A. (2015). Lexical stress in English pronunciation. In M. Reed & J. M. Levis (Eds.), *The handbook of English pronunciation* (1st ed., pp. 106–124). Wiley.

Darcy, I., Peperkamp, S., & Dupoux, E. (2007). Bilinguals play by the rules: Perceptual compensation for assimilation in late L2-learners. In J. Cole & J. Hualde (Eds.), *Laboratory Phonology 9* (pp. 411–442). Mouton de Gruyter.

Darcy, I., Ramus, F., Christophe, A., Kinzler, K., & Emmanuel, D. (2009). Phonological knowledge in compensation for native and non-native assimilation. In F. Kügler, C. Féry, & R. V. D. Vijver (Eds.), *Variation and gradience in phonetics and phonology* (pp. 265–310). Mouton de Gruyter.

Decoo, W. (1996). The induction-deduction opposition: Ambiguities and complexities of the didactic reality. *International Review of Applied Linguistics in Language Teaching, 34*(2), 95–118. https://doi.org/10.1515/iral.1996.34.2.95

Derwing, T. M., & Rossiter, M. J. (2003). The effects of pronunciation instruction on the accuracy, fluency, and complexity of L2 accented speech. *Applied Language Learning, 13*(1), 1–17.

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: evidence-based perspectives for L2 teaching and research* (1st ed.). John Benjamins Publishing Company.

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference.* Palgrave Macmillan.

Ellis, R. (2009). Implicit and explicit learning, knowledge and instruction. In R. Ellis (Ed.), *Implicit and explicit knowledge in second language learning, testing and teaching* (pp. 3–25). Multilingual Matters.

Escudero, P. (2001). The role of the input in the development of L1 and L2 sound contrasts: Language-specific cue weighting for vowels. In A. H.-J. Do, L. Domínguez, & A. Johansen (Eds.), *Proceedings of the 25th annual Boston University Conference on Language Development*, (pp. 250–261). Cascadilla Press.

Escudero, P. (2005). *Linguistic perception and second language acquisition: Explaining the attainment of optimal phonological categorization*. [Doctoral dissertation, University of Utrecht].

Escudero, P. (2007). Second language phonology: The role of perception. In M. C. Pennington (Ed.), *Phonology in context* (pp. 109–134). Palgrave Macmillan.

Escudero P., & Boersma, P. (2002). The subset problem in L2 perceptual development: Multiple-category assimilation by Dutch learners of Spanish. In B. Skarabela, S. Fish, & A. H.-J. Do (Eds.), *Proceedings of the 26th annual Boston University Conference on Language Development* (pp. 208–219). Cascadilla Press.

Escudero, P., & Boersma, P. (2004). Bridging the gap between L2 speech perception research and phonological theory. *Studies in Second Language Acquisition*, *26*(4), 551–585.

ETS. (2025). *Comparing TOEFL iBT scores to CEFR levels*. https://www.ets.org/toefl/institutions/ibt/compare-scores.html

ETS Global. (2024). *The Common European Framework of Reference for Languages (CEFR)*. https://www.etsglobal.org/fr/en/content/common-european-framework-reference-languages

Field, J. (1999). Key concept: Bottom up versus top down. *ELT Journal, 53*, 338–339.

Field, J. (2003). Promoting perception: Lexical segmentation in L2 listening. *ELT Journal*, *57*(4), 325–334.

Field, J. (2008a). Bricks or mortar: Which parts of the input does a second language listener rely on? *TESOL Quarterly*, *42*(3), 411–432.

Field, J. (2008b). Emergent and divergent: A view of second language listening research. *System, 36*(1), 2–9. https://doi.org/10.1016/j.system.2008.01.001

Fisher, R. A. (1921). On the “probable error” of a coefficient of correlation deduced

from a small sample. *Metron, 1*(4), 3–32.

Flege, J. E. (1987). The production of "new" and "similar" phones in a foreign language: Evidence for the effect of equivalence classification. *Journal of Phonetics, 15*, 47–65.

Flege, J. E. (1995). Second language speech learning: Theory, findings and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233–277). York Press.

Flege, J. E. (2003). Assessing constraints on second-language segmental production and perception. In A. Meyer & Schiller, N. (Eds.), *Phonetics and phonology in language comprehension and production: Differences and similarities* (pp. 319–355). Mouton de Gruyter.

Flege, J. E., & Bohn, O.-S. (2021). The revised speech learning model (SLM-r). In R. Wayland (Ed.), *Second language speech learning* (pp. 3–83). Cambridge University Press. https://doi.org/10.1017/9781108886901.002

Flege, J. E., & Wang, C. (1989). Native-language phonotactic constraints affect how well Chinese subjects perceive the word-final English /t/-/d/ contrast. *Journal of Phonetics*, *17*(4), 299–315.

Flege, J. E., Bohn, O.-S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics, 25*(4), 437–470.

Flowerdew, J., & Miller, L. (2005). *Second language listening: Theory and practice*. Cambridge University Press.

Foote, J. A., & Trofimovich, P. (2018). Is it because of my language background? A study of language background influence on comprehensibility judgments. *Canadian Modern Language Review*, *74*(2), 253–278. https://doi.org/10.3138/cmlr.2017-0011

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct–realist perspective. *Journal of Phonetics*, *14*(1), 3–28.

Giannakopoulou, A., Brown, H., Clayards, M., & Wonnacott, E. (2017). High or low? Comparing high and low-variability phonetic training in adult and child second language learners. *PeerJ*, *5*, e3209. https://doi.org/10.7717/peerj.3209

Gilbert, J. (1995). Pronunciation practice as an aid to listening comprehension. In J. Mendelsohn & J. Rubin (Eds.), *A guide for the teaching of second language listening* (pp. 97–112).

Gilbert, J. B. (2019). An informal account of how I learned about English rhythm. *TESOL Journal, 10,* Article e441. https://doi.org/10.1002/tesj.441

GOV.UK. (2022). Guidance: Prove your English language abilities with a secure English language test (SELT). *Citizenship.* Retrieved 26 June, 2022 from https://www.gov.uk/guidance/prove-your-english-language-abilities-with-a-secure-english-language-test-selt

Grabe, E., Rosner, B. S., García-Albea, J. E., & Xiaolin Zhou. (2003). Perception of English intonation by English, Spanish, and Chinese listeners. *Language and Speech*, *46*(4), 375–401.

Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2020). bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software, 92*(10), 1–29. doi.org/10.18637/jss.v092.i10

Hawkins, S., & Smith, R. (2001) Polysp: A polysystemic, phonetically-rich approach to speech understanding. *Italian Journal of Linguistics 13*(1), 99–189.

Henrichsen, L. E. (1984). Sandhi-variation: a filter of input for learners of ESL. *Language Learning, 34*(3), 103–123.

Hieke, A. E. (1984). Linking as a marker of fluent speech. *Language and Speech, 27,*

343–354.

Hieke, A. E. (1987a). Absorption and fluency in native and non-native casual speech in English. In A. James & J. Leather (Eds.), *Sound patterns in second language acquisition* (pp. 41–58). Foris.

Hieke, A. E. (1987b). The resolution of dynamic speech in L2 listening. *Language Learning, 37(*1), 123–140.

Hulstijn, J. H. (2003). Connectionist models of language processing and the training of listening skills with the aid of multimedia software. *Computer Assisted Language Learning*, *16*(5), 413–425.

IELTS. (2025). *Compare IELTS*. https://ielts.org/organisations/ielts-for-organisations/compare-ielts

International Phonetic Association. (2020). *The International Phonetic Alphabet and the IPA chart*. https://www.internationalphoneticassociation.org/content/ipa-chart

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly, 10*(2), 135–159. https://doi.org/10.1080/15434303.2013.769545

Ito, Y. (2001). Effect of reduced forms on ESL learners' input-intake process. *Second Language Studies*, *20*(1), 99–124.

Ito, Y. (2006). The significance of reduced forms in L2 pedagogy. In J. D. Brown & K. Kondo-Brown (Eds.), *Perspectives on teaching connected speech to second language speakers* (pp. 17–25). National Foreign Language Resource Center, University of Hawai'i at Manoa.

Iverson, P., & Evans, B. G. (2007). Learning English vowels with different first-language vowel systems: Perception of formant targets, formant movement, and

duration. *Journal of the Acoustical Society of America, 122*(5), 2842–2854. https://doi.org/10.1121/1.2783198

Iverson, P., & Evans, B. G. (2009). Learning English vowels with different first-language vowel systems II: Auditory training for native Spanish and German speakers. *Journal of the Acoustical Society of America, 126*(2), 866–877.

Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English/r/-/l/to Japanese adults. *The Journal of the Acoustical Society of America, 118*(5), 3267–3278.

Jeffreys, H. (1961). *Theory of Probability* (3rd ed.). Oxford University Press.

Jones, K., & Ono, T. (2001) Reconciling textbook and naturally occurring talk: What we think we do is not what we do. *Journal of Second Language Acquisition and Teaching, 8*, 1–13.

Kennedy, S., & Trofimovich, P. (2008). Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context. *Canadian Modern Language Review*, *64*(3), 459–489. https://doi.org/10.3138/cmlr.64.3.459

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics, 50*, 93–107.

Kuhl, P. K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences USA, 97*, 11850–11857.

Kul, M. (2016). Effects of two teaching methods of connected speech in a Polish EFL classroom. *Research in Language*, *14*(4), 389–407.

Kuo, F., Kuo, Y., & Lee, J. (2016). Effects of communicative instruction versus explicit instruction on Taiwanese EFL junior high students' word recognition of connected

speech. *International Journal of Language and Linguistics*, *3*(2), 101–111.

Ladefoged, P., & Johnson, K. (2015). *A course in phonetics* (7th ed.). Cengage Learning.

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, *44*(2), 325–343.

Lemhöfer, K., & Broersma, M. (n.d.). What is LexTALE? *LexTALE* https://www.lextale.com/whatislextale.html

Levelt, W. J. M. (1993). Language use in normal speakers and its disorders. In G. Blanken, J. Dittmann, H. Grimm, J. C. Marshall & C.-W. Wallesch (Eds.), *Linguistic disorders and pathologies* (pp. 1–15). De Gruyter.

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

Levy, E. S. (2009a). Language experience and consonantal context effects on perceptual assimilation of French vowels by American-English learners of French. *The Journal of the Acoustical Society of America, 125*(2), 1138–1152. https://doi.org/10.1121/1.3050256

Levy, E. S. (2009b). On the assimilation-discrimination relationship in American English adults' French vowel learning. *The Journal of the Acoustical Society of America, 126*(5), 2670–2682. https://doi.org/10.1121/1.3224715

Liang, D. (2015). Chinese learners' pronunciation problems and listening difficulties in English connected speech. *Asian Social Science, 11*(16), 98–106. https://doi.org/10.5539/ass.v11n16p98

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability

in learning new perceptual categories. *The Journal of the Acoustical Society of America*, *94*, 1242–1255. https://doi.org/10.1121/1.408177

Lively, S. E., Pisoni, D. B., Yamada, R. A., Tohkura, Y., & Yamada, T. (1994). Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of new phonetic categories. *The Journal of the Acoustical Society of America, 96*(4), 2076–2087. https://doi.org/10.1121/1.410149

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English/r/ and /l/ : a first report. *The Journal of the Acoustical Society of America, 89*(2), 874–886. https://doi.org/10.1121/1.1894649

Ludwig, A., & Mora, J. C. (2017). Processing time and comprehensibility judgments in non-native listeners' perception of L2 speech. *Journal of Second Language Pronunciation, 3*(2), 167–198. https://doi.org/10.1075/jslp.3.2.01lud

Lynch, T. (1998). Theoretical perspectives on listening. *Annual Review of Applied Linguistics*, *18*, 3–19.

Lynch, T. (2006). Academic listening: Marrying top and bottom. In E. Usó-Juan & A. Martínez-Flor (Eds.), C*urrent trends in the development and teaching of the four language skills* (pp. 91–110). Mouton de Gruyter.

Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition, 19*(1), 37–66. https://www.jstor.org/stable/44488666

Magnuson, J. S. (2017). Mapping spoken words to meaning. In G. Gaskell, & J. Mirković (Eds.), *Speech perception and spoken word recognition* (pp. 76–96). Psychology Press.

Major, R. C. (2001). *Foreign accent: The ontogeny and phylogeny of second language phonology.* Erlbaum.

Major, R. C. (2002). The phonology of the L2 user. In V. Cook (Ed.), *Portraits of the L2 user* (pp. 65–92). Multilingual Matters.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*(1), 29–63.

Marsman, M., & Wagenmakers, E. J. (2017). Three insights from a Bayesian interpretation of the one–sided P value. *Educational and Psychological Measurement, 77*(3), 529–539. https://doi.org/10.1177/0013164416669201

Martínez-Flor, A., & Usó-Juan, E. (2006). Towards acquiring communicative competence in listening. In E. Usó-Juan and A. Martínez-Flor (Eds.), C*urrent trends in the development and teaching of the four language skills* (pp. 29–46). Mouton de Gruyter.

Matsuzawa, T. (2006). Comprehension of English reduced forms by Japanese business people and the effectiveness of instruction. In J. D. Brown & K. Kondo-Brown, (Eds.), *Perspectives on teaching connected speech to second language speakers* (pp. 59–66). National Foreign Language Resource Center, University of Hawai'i at Manoa.

Mendelsohn, D. (2006). Learning how to listen using learning strategies. In E. Usó-Juan & A. Martínez-Flor (Eds.), *Current trends in the development and teaching of the four language skills* (pp. 75–90). Mouton de Gruyter. https://doi.org/10.1515/9783110197778.2.75

Michaels, D. (1974). Sound replacements and phonological systems. *Linguistics*, *12*(126), 69–81.

Ministry of Labor, Executive Yuan, Taiwan. (n.d.). *The history of enacting and adjusting the minimum wage policy*. Ministry of Labor, Executive Yuan, Taiwan. Retrieved July 9, 2025, from

https://english.mol.gov.tw/21004/21015/21016/21033/21232/

Momen, N., & Pilus, Z. (2022). The effects of connected speech instruction on second or foreign language learners' perceptive skills and connected speech production: A systematic review of the literature (2000–2021). *International Journal of Learning, Teaching and Educational Research, 21*(9), 389–414.

Mora, J. C. (2022). Accentedness and comprehensibility in non-native listeners' perception of L2 speech. In V. G. Sardegna & A. Jarosz (Eds.), *Theoretical and practical developments in English speech assessment, research, and training* (pp. 109–125). Springer. https://link.springer.com/chapter/10.1007/978-3-030-98218-8_7

Mora, J. C., Ortega, M., Mora-Plaza, I., & Aliaga-García, C. (2022). Training the pronunciation of L2 vowels under different conditions: The use of non-lexical materials and masking noise. *Phonetica, 79*(1), 1–43. https://doi.org/10.1515/phon-2022-2018

Mora-Plaza, I., Ortega, M., & Mora, J. C. (2022). High-variability phonetic training under different conditions: Individual differences in auditory attention control. In V. G. Sardegna & A. Jarosz (Eds.), *Theoretical and practical developments in English speech assessment, research, and training* (pp. 241–260). Springer. https://doi.org/10.1007/978-3-030-98218-8_14

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, *76*(2), 165–178.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45*(1), 73–97.

Munro, M. J., & Derwing, T. M. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, *19*(1), 1–16.

Nagle, C. (2018). Motivation, comprehensibility, and accentedness in L2 Spanish: investigating motivation as a time-varying predictor of pronunciation development. *The Modern Language Journal, 102*, 199–217. https://doi.org/10.1111/modl.12461

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47*(2), 204–238. https://doi.org/10.1016/S0010-0285(03)00006-9

Norris, J., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning, 50*, 417–528.

Norris, R. (1995). Teaching reduced forms: Putting the horse before the cart. *English Teaching Forum, 33*, 47–50.

Office for Standards in Education (Ofsted). (2021). *Curriculum research review series: Languages*. https://www.gov.uk/government/publications/curriculum-research-review-series-languages/curriculum-research-review-series-languages

Ortega, M., Mora, J. C., & Mora-Plaza, I. (2022). L2 learners' self-assessment of comprehensibility and accentedness: Over/under-estimation, effects of rating peers, and attention to speech features. In J. Levis & A. Guskaroska (eds.), *Proceedings of the 12th Pronunciation in Second Language Learning and Teaching Conference*, held June 2021 virtually at Brock University, St. Catharines, ON. https://doi.org/10.31274/psllt.13354.

Osada, N. (2004). Listening comprehension research: A brief review of the past thirty years. *Dialogue*, *3*, 53–66.

Pearson. (2024). *PTE Academic: Test taker score guide* (Version 21). Pearson. https://www.pearsonpte.com/ctf-assets/yqwtwibiobs4/5Sz9Ur4qbus8AEOQdetkAj/69f6c1f2e2870980740b10a2ea9b467f/pte-academic-test-taker-score-guide-nov-2024-v4.pdf

Rahimi, M., & Chalak, A. (2017). The effect of connected speech teaching on listening comprehension of Iranian EFL learners. *Journal of Applied Linguistics and Language Research*, *4*(8), 280–291.

Rankin, P. T. (1928). The importance of listening ability. *The English Journal*, *17*(8), 623–630.

Reed, M., & Michaud, C. (2011). An integrated approach to pronunciation: listening comprehension and intelligibility in theory and practice. In J. Levis & K. LaVelle (Eds.), *Proceedings of the pronunciation in second language learning and teaching conference* (pp. 95–104). Iowa State University.

Renandya, W. A., & Farrell, T. S. C. (2011). 'Teacher, the tape is too fast!' Extensive listening in ELT. *ELT Journal, 65*(1), 52–59. https://doi.org/10.1093/elt/ccq015

Rost, M. (2001). Listening. In R. Carter & D. Nunan (Eds.), *The Cambridge guide to teaching English to speakers of other languages* (pp. 7–13). Cambridge University Press.

Rumelhart, D. E., & Ortony, A. (1977). The representation of knowledge in memory. In R. C. Anderson, R. J. Spiro & W. E. Montague (Eds.), *Schooling and the acquisition of knowledge* (pp. 99–135). Erlbaum.

Rumelhart, D. E. (1980). Schemata: The building blocks of cognition. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical Issues in Reading Comprehension: Perspectives from cognitive psychology, linguistics, artificial intelligence and education* (pp. 33–58). Routledge. https://doi.org/10.4324/9781315107493-4

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

Saito, K., Dewaele, J.-M., Abe, M., & In'nami, Y. (2018). Motivation, emotion, learning experience, and second language comprehensibility development in classroom settings: a cross-sectional and longitudinal study. *Language Learning, 68*, 709–743. https://doi.org/10.1111/lang.12297

Saito, K., Tran, M., Suzukida, Y., Sun, H., Magne, V., & Ilkan, M. (2019). How do second language listeners perceive the comprehensibility of foreign-accented speech?: Roles of first language profiles, second language proficiency, age, experience, familiarity, and metacognition. *Studies in Second Language Acquisition*, *41*(5), 1133–1149. https://doi.org/10.1017/S0272263119000226

Samuel, A. G. (1981). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General, 110*(4), 474–494. https://doi.org/10.1037/0096-3445.110.4.474

Schad, D. J., Nicenboim, B., Bürkner, P.-C., Betancourt, M., Vasishth, S., & Steinley, D. (2023). Workflow techniques for the robust use of Bayes factors. *Psychological Methods, 28*(6), 1404–1426. https://doi.org/10.1037/met0000472

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*(2), 129–158. https://doi.org/10.1093/applin/11.2.129

Sheerin, S. (1987). Listening comprehension: Teaching or testing? *ELT Journal*, *41*(2), 126–131.

Shockey, L. (2003). *Sound Patterns of Spoken English*. Blackwell Publishing.

Siegel, J. (2018). Teaching bottom-up and top-down strategies. In J. I. Liontas, T. International Association, & M. DelliCarpini (Eds.), *The TESOL encyclopedia of English language teaching*. https://doi.org/10.1002/9781118784235.eelt0597

Silvey, C., Dienes, Z., & Wonnacott, E. (2024). Bayes factors for logistic (mixed effect) models. https://doi.org/10.31234/osf.io/m4hju

Strange, W., & Shafer, V. L. (2008). Speech perception in second language learners: The re-education of selective perception. In J. G. Hansen Edwards & M. L. Zampini (Eds.), *Studies in bilingualism* (Vol. 36, pp. 153–191). John Benjamins Publishing Company.

Taylor Reid, K., Trofimovich, P., & O'Brien, M. G. (2019). Social attitudes and speech ratings: Effects of positive and negative bias on multiage listeners' judgments of second language speech. *Studies in Second Language Acquisition, 41*, 419–442.

Taylor Reid, K., Trofimovich, P., O'Brien, M. G., & Tsunemoto, A. (2022). Using task practice to reduce social influences on listener evaluations of second language accent and comprehensibility. *International Journal of Listening, 36*(3), 283–298. https://doi.org/10.1080/10904018.2021.1904933

Teimouri, Y., Plonsky, L., & Tabandeh, F. (2020). L2 grit: Passion and perseverance for second-language learning. *Language Teaching Research*, *26*(5), 893–918. https://doi.org/10.1177/1362168820921895

Ting, W. Y., & Kuo, F. L. (2012). Messages behind the unheard sounds: Crossing the word boundaries through songs. *NCUE Journal of Humanities, 5*, 75–92.

Trofimovich, P., Isaacs, T., Kennedy, S., Saito, K., & Crowther, D. (2016). Flawed self-assessment: Investigating self- and other-perception of second language speech. *Bilingualism: Language and Cognition, 19*(1), 122–140.

Trofimovich, P., Nagle, C. L., O'Brien, M. G., Kennedy, S., Taylor Reid, K., & Strachan, L. (2020). Second language comprehensibility as a dynamic construct. *Journal of Second Language Pronunciation, 6*, 430–457. https://doi.org/10.1075/jslp.20003.tro

Trofimovich, P., Tekin, O., & Lindberg, R. (2024). Listening and comprehensibility. In E. Wagner, A. O. Batty, & E. Galaczi (Eds.), *The Routledge Handbook of Second*

Language Acquisition and Listening (pp. 201–213). Routledge. https://doi.org/10.4324/9781003219552-17

Tsui, A. B. M., & Fullilove, J. (1998). Bottom-up or top-down processing as a discriminator of L2 listening performance. *Applied Linguistics, 19*(4), 432–451. https://doi.org/10.1093/applin/19.4.432

van Leussen, J.-W., & Escudero, P. (2015). Learning to perceive and recognize a second language: the L2LP model revised. *Frontiers in Psychology*, *6*, Article 1000. https://doi.org/10.3389/fpsyg.2015.01000

Vandergrift, L. (2004). Listening to learn or learning to listen? *Annual Review of Applied Linguistics*, *24*, 3–25.

Vandergrift, L., & Goh, C. C. M. (2012). *Teaching and learning second language listening: Metacognition in action*. Routledge.

Vidal, K. (2019). Perception of phonological assimilation and elision by ESL learners and its impact on listening proficiency. *IUP Journal of English Studies*, *14*(1), 93–111.

Walker, N. (2014). Listening: The most difficult skill to teach. *Encuentro*, *23*, 167–175.

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

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science, 167*(3917), 392–393. https://doi.org/10.1126/science.167.3917.392

Weinstein, N. (2001). *Whaddaya say?: Guided practice in relaxed speech* (2nd ed.). Longman.

Wilson, M. (2003). Discovery listening—Improving perceptual processing. *ELT Journal*, *57*(4), 335–343.

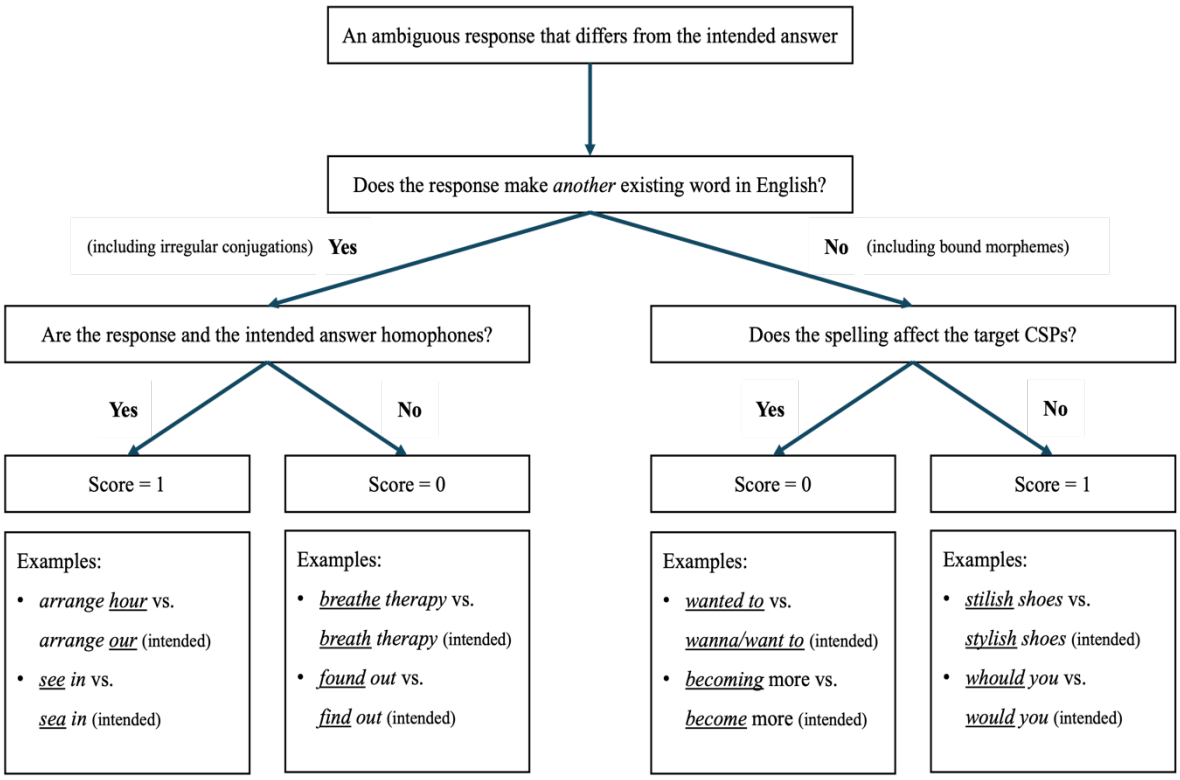
Wong, S. W. L., Mok, P. P. K., Chung, K. K.-H., Leung, V. W. H., Bishop, D. V. M., &

Chow, B. W.-Y. (2017). Perception of native English reduced forms in Chinese learners: Its role in listening comprehension and its phonological correlates. *TESOL Quarterly*, *51*(1), 7–31.

YouGlish. (n.d.). *YouGlish*. https://youglish.com/

Appendix A

Scoring Criteria for Ambiguous Responses or Typographical Errors in Dictation Measures



Appendix B

CEFR Level Equivalents across Common Standardised English Proficiency Tests

Table B1 below summarises the CEFR-equivalent score ranges for four widely used standardised English proficiency tests: IELTS, TOEFL iBT, TOEIC (Reading/Listening), and PTE Academic.

CEFR Level		IELTS	TOEFL iBT	TOEIC (Reading/Listening)	PTE Academic
Proficient User	C2	8.5 – 9.0	114 – 120	n/a	85 – 90
	C1	7.0 – 8.0	95 – 113	945 – 990	76 – 84
Independent User	B2	5.5 – 6.5	72 – 94	785 – 940	59 – 75
	B1	4.5 – 5.0	42 – 71	550 – 780	43 – 58
Basic User	A2	n/a	n/a	225 – 545	30 – 42
	A1	n/a	n/a	120 – 220	10 – 29

Table B1. CEFR level equivalents across common standardised English proficiency tests.

Note. CEFR equivalents are based on guidelines published by IELTS, ETS (TOEFL and TOEIC), and Pearson (PTE Academic). See References for full source details.

Appendix C

Complete Stimulus Inventory Developed for the Study

To maintain formatting and legibility, the complete stimulus inventory is provided as a separate spreadsheet, available at: [Appendix C Stimulus Inventory].

The materials included are provided for academic transparency and review. They may not be reproduced, distributed, or used in other research or publications without the express permission of the author.

Appendix D

Audio Recording Procedures and Instructions for Contributors

- **Audio Contributors**

  Four native speakers of UK English were recruited to provide the recordings, comprising two male and two female speakers. Each contributor completed an estimated 30 minutes of recording.

- **Stimulus Assignment**

  Stimuli were assigned to contributors through a balanced random selection process to ensure an even distribution across CSP categories. Table D1 summarises the number of experimental stimuli, target words, and distractor items allocated to each speaker[40].

Contributor	Exp. Stimuli	Target Words	Distractors	Total Items
Speaker 1 (M)	29	60	14	43 sentences + 60 words
Speaker 2 (M)	29	59	15	44 sentences + 59 words
Speaker 3 (F)	29	58	15	44 sentences + 58 words
Speaker 4 (F)	29	58	14	43 sentences + 58 words
Total	116 sentences	235 words	58 sentences	174 sentences + 235 words

Table D1. Stimuli assigned to each audio recording contributor.

Note. Exp. = Experimental; M = Male; F = Female

- **Recording Procedures**

Part 1: Sentence Recording

Instructions to Contributors:

[40]These were the four initial contributors. Following phonetic analysis, a subset of stimuli requiring a second recording was produced by two additional speakers (one male, one female), following the same procedures.

Read each sentence AS NATURALLY AS POSSIBLE. The purpose is to capture everyday spoken English, NOT 'dictionary English'. You can imagine talking with a close friend who is also a native English speaker.

- **Step 1:** Silently read 7 – 8 experimental stimuli and 3 – 4 distractors per round.
- **Step 2:** Record each sentence three times, with a 2-second pause between repetitions.
- [Repeat Steps 1 – 2 for four rounds, yielding a total of 43 – 44 recoded sentences per contributor]

Part 2: Word Recording

Instructions to Contributors:

Pronounce each word AS CLEARLY AS POSSIBLE. The purpose is to capture the pronunciation of words in isolation (context is to be disregarded). You can imagine teaching individual words to someone learning English.

- **Step 1:** Silently read 14 – 15 sets of target words per round.
- **Step 2:** Record each set of words three times, with a 2-second pause between repetitions.
- [Repeat Steps 1 – 2 for two rounds, yielding a total of 58 – 60 recorded words per speaker.]

- **Recording Equipment and Settings**
  - **Microphone:** SUDOTACK ST-800 condenser microphone

    Frequency Response: 30Hz – 16kHz

    Accessories: foam windscreen, shock mount, pop filter
  - **Sampling Rate:** 48kHz
  - **Bit Depth:** 32-bit float
  - **Environment:** a soundproof laboratory at the University of Oxford

Appendix E

Instructional Materials for Individual Intervention Conditions

To maintain formatting and legibility, the instructional materials for all three intervention conditions are provided as separate files, available at: [Appendix E Instructional Materials].

The materials included are provided for academic transparency and review. They may not be reproduced, distributed, or used in other research or publications without the express permission of the author.

Appendix F

Phase 2 LexTALE and LSK Analyses

LexTALE

The LexTALE score was utilised as a proxy for general English proficiency to ensure participants across the three intervention groups exhibited comparable proficiency levels at the outset of the experiment. This measure provided a supplement to participants' self-reported proficiency and the inclusion criteria (outlined in Section 5.1.1). Figure F1 presents the mean LexTALE scores by group along with 95% confidence intervals.

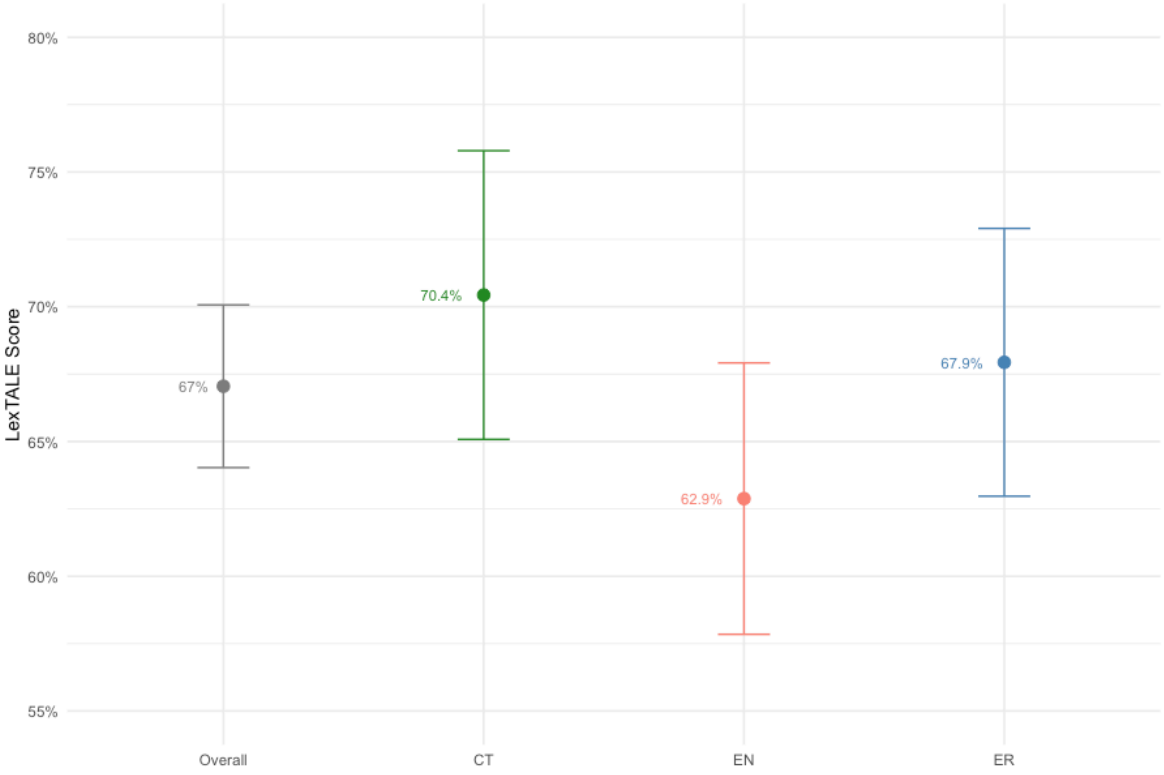


Figure F1. LexTALE Sores by Group (at Pre-test).

While all participants possessed a proficiency level of B2 or higher and were randomly assigned to the intervention groups, variability in group means was observed. The CT group exhibited the highest mean score (70.4%), followed by the ER group (67.9%),

while the EN group displayed the lowest mean (62.9%). To assess the degree of the certainty surrounding these group means, a Bayesian regression model was fitted, with two versions capturing all pairwise contrasts (see Section 5.2.1 for details on the simple coding scheme). The model was run with four chains of 1,500 MCMC sampling iterations (totalling 6,000), excluding 500 warm-ups per chain. The results are summarised in Table F1.

Model for LexTALE: Version 1

$$LexTALE\ score \sim CT\_VERSUS\_ER\ +\ CT\_VERSUS\_EN$$

Model for LexTALE: Version 2

$$LexTALE\ score \sim ER\_VERSUS\_CT\ +\ ER\_VERSUS\_EN$$

	Estimate	**Est. Error**	**CI_lower**	**CI_upper**	**B<>0**
Intercept	0.67	0.02	0.64	0.70	1
CT_VERSUS_ER	-0.02	0.04	-0.10	0.05	0.75
CT_VERSUS_EN	-0.08	0.04	-0.15	0.00	0.98
ER_VERSUS_EN	-0.05	0.04	-0.13	0.03	0.91

Table F1. Summary of Bayesian Model for LexTALE.

The intercept estimate corresponds to an average LexTALE score of 67% accuracy, with the CIs [0.64, 0.70] showing a high degree of certainty. Among the group contrasts, the largest difference is observed between CT and EN Groups (b = -0.08, CIs[-0.15, 0.00]) with the upper boundary of the CIs situated at zero, providing evidence of a meaningful difference in proficiency between these two groups. This is further reflected in a B<>0

value of 0.98, indicating that 98% of 6,000 sampling iterations yield a negative estimate for CT_VERSUS_EN. To account for this baseline difference, LexTALE scores were included as a covariate in the models for principal measures (as presented in the corresponding sections), ensuring that the intervention effects could be assessed while controlling for variations in initial proficiency levels. In contrast, the CIs for the other two contrasts cross zero (CT_VERSUS_ER: b = -0.02, CIs[-0.10, 0.05]; ER_VERSUS_EN: b = -0.05, CIs[-0.13, 0.03]), indicating no evidence of meaningful differences between the CT and ER groups or between the ER and EN groups.

L1 Lexical-Semantic Knowledge Test (LSK)

The LSK was administered to examine whether there were differences across intervention groups in their semantic knowledge of target words – i.e. whether participants understood the meanings of target words within the context of sentential stimuli when visually presented. As outlined in the methodology, since the vocabulary items were deliberately controlled to exclude advanced words (cf. Table 3.2), participants were expected to recognise the majority of the items. Figure F2 presents the mean LSK scores by group and assessment along with 95% confidence intervals.

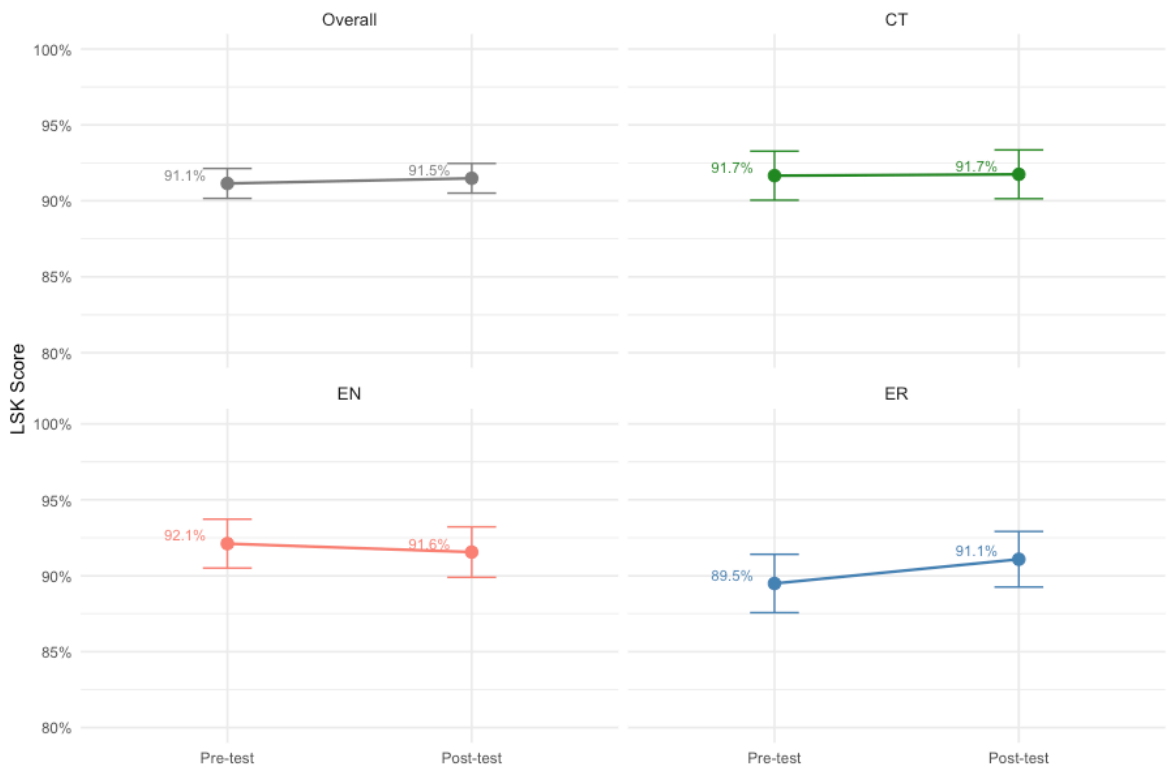


Figure F2. LSK Scores by Group and Assessment.

As depicted in Figure F2, the overall mean accuracy exceeded 91% at both pre-test and post-test, with only a minimal difference of 0.4 percentage points. For individual groups, accuracy remained consistently high, with slight variations across assessments: the CT group demonstrates identical accuracy at the pre-test and post-test, the EN group a slight decrease of 0.5 percentage points, and the ER group an increase of 1.6 percentage points. In line with Phase 1 results, the consistent and high accuracy indicate that the vast majority of target lexical items were familiar to participants and validate the deliberate control over the vocabulary level when devising the stimuli, which ensured that advanced (C1 or above) words were excluded. Consequently, participants in both phases exhibited similar levels of familiarity with the target words, despite a slightly different proficiency requirement applied (B2 for Phase 2 and C1 for Phase 1) to facilitate participant recruitment.

While neither overall changes nor group differences were hypothesised, a Bayesian logistic mixed effects model (two versions for pairwise comparisons) was fitted for statistical examination, with specifications similar to those applied to other measures. The model was run with four chains of 1,500 iterations, excluding 500 warm-ups per chain. Table F2 summarises the results of the Bayesian model.

Model for LSK: Version 1

$$LSK\ score \sim CT\_VERSUS\_ER * Assessment.ct + CT\_VERSUS\_EN * Assessment.ct + (Assessment.ct \mid PID)$$

Model for LSK: Version 2

$$LSK\ score \sim ER\_VERSUS\_CT * Assessment.ct + ER\_VERSUS\_EN * Assessment.ct + (Assessment.ct \mid PID)$$

	Estimate	**Est. Error**	**CI_lower**	**CI_upper**	**B<>0**
Intercept	2.42	0.07	2.28	2.57	1
Assessment.ct	0.03	0.10	-0.16	0.23	0.63
Assessment.ct (CT)	0.01	0.17	-0.32	0.34	0.52
Assessment.ct (ER)	0.16	0.19	-0.23	0.54	0.80
Assessment.ct (EN)	-0.06	0.19	-0.43	0.31	0.64
CT_VERSUS_ER:Assessment.ct	0.16	0.23	-0.30	0.60	0.75
CT_VERSUS_EN:Assessment.ct	-0.09	0.23	-0.55	0.37	0.65
ER_VERSUS_EN:Assessment.ct	-0.24	0.23	-0.70	0.21	0.85

Table F2. Summary of Bayesian Model for LSK.

Note. Estimates are in log-odds space.

An estimate of 2.42 for the intercept corresponds to the consistently high overall accuracy of approximately 91%, with the CIs[2.28, 2.57] indicating substantial statistical certainty. As expected, the main effect (Assessment.ct: b = 0.03, CIs[-0.16, 0.23]) shows no evidence of a meaningful difference from the pre-test to post-test, as the CIs cross zero. Similarly, all group-specific effects yield estimates with highly uncertain CIs: CT group (b = 0.01, CIs[-0.32, 0.34]); ER group (b = 0.16, CIs[-0.23, 0.54]); EN group (b = -0.06, CIs[-0.43, 0.31]). Furthermore, the CIs for CT_VERSUS_ER:Assessment.ct (b = 0.16, CIs[-0.30, 0.60]), CT_VERSUS_EN:Assessment.ct (b = -0.09, CIs[-0.55, 0.37]), and ER_VERSUS_EN:Assessment.ct (b = -0.24, CIs[-0.70, 0.21]) again all cross zero, indicating a lack of evidence of any meaningful interactions. Thus, these results reaffirm participants' high level of familiarity with the target lexical items, which remained consistent across groups and assessments.

Appendix G

Phase 2 Frequentist Mixed Effects Regression Model Outputs for Comparison

These mixed effects models were fitted using the *lme4* package (Bates et al. 2015). The outputs are presented in the order of (1) connected speech dictation, (2) disparity score, (3) comprehensibility rating, and (4) general listening comprehension, corresponding to the sequence in which these measures are reported in the main text. The model syntax is included for reference. For details on the simple coding scheme and the two model versions, refer to Section 5.2.1.

Connected Speech Dictation

Model for Connected Speech: Version 1

$$Connected\ speech \sim CT\_VERSUS\_ER * Assessment.ct + CT\_VERSUS\_EN * Assessment.ct + LexTALE.ct + (Assessment.ct \mid PID)$$

Model for Connected Speech: Version 2

$$Connected\ speech \sim ER\_VERSUS\_CT * Assessment.ct + ER\_VERSUS\_EN * Assessment.ct + LexTALE.ct + (Assessment.ct \mid PID)$$

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
Intercept	-0.03	0.06	-0.42	0.67
LexTALE.ct	3.54	0.53	6.72	1.85e-11 ***
Assessment.ct	**0.17**	**0.09**	**1.81**	**0.04** *
CT_VERSUS_ER	0.05	0.16	0.30	0.76
CT_VERSUS_EN	0.24	0.16	1.55	0.12

ER_VERSUS_EN	0.19	0.16	1.23	0.22
CT_VERSUS_ER:Assessment.ct	**0.33**	**0.23**	**1.43**	**0.08**
CT_VERSUS_EN:Assessment.ct	**0.29**	**0.22**	**1.30**	**0.10**
ER_VERSUS_EN:Assessment.ct	-0.04	0.23	-0.17	0.87

Table G1. Frequentist logistic mixed effects regression model output for connected speech dictation (cf. Table 5.6).

Note. For parameters with directional hypotheses (rows in bold), the *p*-values are one-tailed.

Disparity Score

Model for Disparity Scores: Version 1

$$Disparity\ score \sim CT\_VERSUS\_ER * Assessment.ct + CT\_VERSUS\_EN * Assessment.ct + LexTALE.ct + (Assessment.ct \mid PID)$$

Model for Disparity Scores: Version 2

$$Disparity\ score \sim ER\_VERSUS\_CT * Assessment.ct + ER\_VERSUS\_EN * Assessment.ct + LexTALE.ct + (Assessment.ct \mid PID)$$

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
Intercept	0.45	0.04	10.41	< 2e-16 ***
LexTALE.ct	2.20	0.36	6.05	1.46e-09 ***
Assessment.ct	**0.15**	**0.08**	**1.80**	**0.04** *
CT_VERSUS_ER	0.03	0.11	0.27	0.78
CT_VERSUS_EN	0.20	0.11	1.90	0.06
ER_VERSUS_EN	0.17	0.11	1.60	0.11

CT_VERSUS_ER:Assessment.ct	**0.31**	**0.20**	**1.50**	**0.07**
CT_VERSUS_EN:Assessment.ct	**0.13**	**0.20**	**0.69**	**0.25**
ER_VERSUS_EN:Assessment.ct	-0.17	0.20	-0.84	0.40

Table G2. Frequentist logistic mixed effects regression model output for disparity scores (cf. Table 5.7).

Note. For parameters with directional hypotheses (rows in bold), the *p*-values are one-tailed.

Comprehensibility Rating

Model for Comprehensibility Ratings: Version 1

$$Comprehensibility\ rating \sim CT\_VERSUS\_ER * Assessment.ct + CT\_VERSUS\_EN * Assessment.ct + LexTALE.ct + (Assessment.ct \mid PID)$$

Model for Comprehensibility Ratings: Version 2

$$Comprehensibility\ rating \sim ER\_VERSUS\_CT * Assessment.ct + ER\_VERSUS\_EN * Assessment.ct + LexTALE.ct + (Assessment.ct \mid PID)$$

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
LexTALE.ct	1.57	0.81	1.95	0.05 .
Assessment.ct	-0.24	0.10	-2.63	0.02 *
CT_VERSUS_ER	-0.03	0.24	-0.11	0.91
CT_VERSUS_EN	0.06	0.24	0.24	0.81
ER_VERSUS_EN	0.08	0.24	0.11	0.91

CT_VERSUS_ER:Assessment.ct	0.05	0.25	0.19	0.85
CT_VERSUS_EN:Assessment.ct	0.35	0.24	1.46	0.14
ER_VERSUS_EN:Assessment.ct	0.31	0.25	1.24	0.22

Table G3. Frequentist ordinal mixed effects regression model output for comprehensibility ratings (cf. Table 5.10).

Threshold	**Estimate**	**Std. Error**	**z value**
1 | 2	-3.83	0.14	-27.73
2 | 3	-2.27	0.11	-21.02
3 | 4	-1.15	0.10	-11.35
4 | 5	-0.41	0.10	-4.14
5 | 6	0.73	0.10	7.34
6 | 7	2.05	0.11	19.26

Table G4. Intercepts (threshold parameters) from the frequentist ordinal mixed effects regression model for comprehensibility ratings (cf. Table 5.11).

Note. The estimates represent the cumulative log-odds of receiving a rating at or below each threshold. These values are estimated as part of the ordinal regression model.

L2 General Listening Comprehension

Model for GLC: Version 1

$$GLC\ score \sim CT\_VERSUS\_ER * Assessment.ct + CT\_VERSUS\_EN * Assessment.ct + (Assessment.ct \mid PID)$$

Model for GLC: Version 2

$$GLC\ score \sim ER\_VERSUS\_CT * Assessment.ct + ER\_VERSUS\_EN * Assessment.ct + (Assessment.ct \mid PID)$$

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
Intercept	1.05	0.08	12.52	<2e-16 ***
Assessment.ct	0.24	0.13	1.82	0.07
CT_VERSUS_ER	0.10	0.21	0.50	0.62
CT_VERSUS_EN	-0.09	0.20	-0.47	0.64
ER_VERSUS_EN	-0.20	0.21	-0.94	0.35
CT_VERSUS_ER:Assessment.ct	-0.18	0.33	-0.54	0.59
CT_VERSUS_EN:Assessment.ct	0.03	0.31	0.10	0.92
ER_VERSUS_EN:Assessment.ct	0.21	0.33	0.64	0.52

Table G5. Frequentist logistic mixed effects regression model output for general listening comprehension (cf. Table 5.13).

Appendix H

R Code for Bayes Factor Computation (Dienes, 2008)

This R code is based on Dienes's (2008) calculator, which computes the ratio of the likelihood of the observed data under the theory model (H_1) versus the null (H_0). The code has been visually formatted using the Carbon[41] display tool to enhance readability and clarity.

```
Bf<-function(sd, obtained, dfdata = 1, likelihood = c("normal", "t"), modeloftheory=
c("normal","t","cauchy", "uniform") ,lower =0, upper=1, modeoftheory = 0, scaleoftheory = 1, dftheory =
1, tail = 2)
{
  if(likelihood=="normal"){
    dfdata=10^10
  }
  if(modeloftheory=="normal"){
    dftheory = 10^10
  } else if(modeloftheory=="cauchy"){
    dftheory = 1
  }
  area <- 0
  normarea <- 0
  if(modeloftheory=="uniform"){
    theta <- lower
    range <- upper - lower
    incr <- range / 2000
    for (A in -1000:1000){
      theta <- theta + incr
      dist_theta <- 1 / range
      height <- dist_theta * dt((obtained-theta)/sd, df=dfdata)
      area <- area + height * incr
    }
    LikelihoodTheory <- area
  }else{
    theta <- modeoftheory - 8 * scaleoftheory
    incr <- scaleoftheory/200
    for (A in -1600:1600){
      theta <- theta + incr
      dist_theta <- dt((theta-modeoftheory)/scaleoftheory, df=dftheory)

      if(tail==1){
        if (theta <= modeoftheory){
          dist_theta <- 0
        } else {
          dist_theta <- dist_theta * 2
        }
      }
      height <- dist_theta * dt((obtained-theta)/sd, df = dfdata)
      area <- area + height * incr
      normarea <- normarea + dist_theta*incr
    }
    LikelihoodTheory <- area/normarea
  }
  LikelihoodNull <- dt(obtained/sd, df = dfdata)
  BayesFactor <- LikelihoodTheory/LikelihoodNull

  BayesFactor

}
```

[41]Carbon (https://carbon.now.sh) is a website tool for creating and formatting images of source code.

Appendix I

Estimation of Predicted Effects for Interactions between Group Contrasts and Assessment.ct Based on Silvey et al. (2024)

To estimate the predicted effects for interactions between group contrasts and assessment (Assessment.ct), we adopted a motivated maximum approach, following the rationale outlined by Silvey et al. (2024).

Consider any pairwise contrast between two groups, denoted as *group1* and *group2*. Let *group1.d* and *group2.d* represent the respective gains from the pre-test and post-test. With all parameters centred, the main effect (*t*) of each group contrast represents the average of *group1.d* and *group2.d*:

$$t = \frac{group1.d + group2.d}{2} \quad \Rightarrow \quad 2t = group1.d + group2.d$$

To derive the maximum possible interaction effect, we consider the case where only one group exhibits improvement (e.g. *group1*), while the other shows no improvement (i.e. *group2.d* = 0). Substituting:

$$2t = group1.d + 0 \quad \Rightarrow \quad group1.d = 2t$$

The interaction effect is defined as the difference in gains between the two groups:

$$interaction = group1.d - group2.d$$

Under the maximum interaction (*mi*) assumption (*group2.d* = 0), this becomes:

$$mi = group1.d - 0 = 2t$$

A predicted interaction (*pi*) can then be derived backwards by halving this value:

$$pi = \frac{mi}{2} = t$$

Thus, for each group contrast, the main effect (*t*) is used as the predicted interaction (*pi*) for the computation of Bayes factors.

Appendix J

Summary of Bayesian Model Output for Disparity Scores by Group and CSP Category

Table J1. below summarises the overall and group-specific effect estimates for individual CSP categories. Separate Bayesian logistic mixed effects models were fitted for each category and each group, with Assessment.ct as the predictor and LexTALE.ct included as a covariate. A visual representation of these results is provided in Figure 5.25, and a tentative classification is discussed in Section 6.2.

CSP Category: *Deletion of Onset*

Main Effect Parameter	Estimate	Est. Error	CI_lower	CI_upper
Assessment.ct (Overall)	0.22	0.16	-0.05	0.49
Assessment.ct (CT)	0.08	0.29	-0.39	0.55
Assessment.ct (ER)	0.59	0.33	0.05	1.14
Assessment.ct (EN)	0.01	0.30	-0.47	0.48

CSP Category: *Glide Insertion*

Main Effect Parameter	Estimate	Est. Error	CI_lower	CI_upper
Assessment.ct (Overall)	-0.02	0.29	-0.48	0.45
Assessment.ct (CT)	-0.41	0.58	-1.36	0.52
Assessment.ct (ER)	0.19	0.59	-0.79	1.16
Assessment.ct (EN)	0.17	0.41	-0.49	0.84

CSP Category: *Reduction of Stop*

Main Effect Parameter	Estimate	Est. Error	CI_lower	CI_upper
Assessment.ct (Overall)	0.23	0.25	-0.17	0.64
Assessment.ct (CT)	0.46	0.47	-0.31	1.24
Assessment.ct (ER)	0.49	0.54	-0.38	1.37
Assessment.ct (EN)	-0.19	0.45	-0.92	0.53

CSP Category: *CC-V Linking*

Main Effect Parameter	Estimate	Est. Error	CI_lower	CI_upper
Assessment.ct (Overall)	0.32	0.17	0.04	0.60
Assessment.ct (CT)	0.09	0.31	-0.41	0.59
Assessment.ct (ER)	0.34	0.35	-0.22	0.91
Assessment.ct (EN)	0.55	0.31	0.06	1.07

CSP Category: *Palatalisation*

Main Effect Parameter	Estimate	Est. Error	CI_lower	CI_upper
Assessment.ct (Overall)	0.22	0.18	-0.08	0.52
Assessment.ct (CT)	0.05	0.32	-0.48	0.58
Assessment.ct (ER)	0.58	0.38	-0.03	1.20
Assessment.ct (EN)	0.03	0.32	-0.50	0.56

CSP Category: *Regressive Assimilation*

Main Effect Parameter	Estimate	Est. Error	CI_lower	CI_upper
Assessment.ct (Overall)	-0.12	0.28	-0.57	0.33
Assessment.ct (CT)	-0.20	0.55	-1.12	0.68

Assessment.ct (ER)	-0.39	0.51	-1.22	0.43
Assessment.ct (EN)	0.12	0.54	-0.75	1.01

Table J1. Summary of Bayesian model output for disparity scores by group and CSP category (cf. Figure 5.25).

Note. All credible intervals (CIs) are reported at the 90% level to reflect directional hypotheses, wherein gains, rather than declines, are expected following targeted interventions.