HAL open science

Linking gestural representations to syllable count judgments: A cross-language test

Anisia Popescu, Ioana Chitoran

To cite this version:

Anisia Popescu, Ioana Chitoran. Linking gestural representations to syllable count judgments: A cross-language test. Laboratory Phonology : Journal of the Association for Laboratory Phonology, 2022, 13 (1), 10.16995/labphon.7681 . hal-03812042

HAL Id: hal-03812042

https://hal.science/hal-03812042v1

Submitted on 11 Feb 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Laboratory Phonology
Journal of the Association for Laboratory Phonology

Popescu, A., & Chitoran, I. 2022. Linking gestural representations to syllable count judgments: A cross-language test. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 13(1): pp. 1–48. DOI: https://doi.org/10.16995/labphon.7681

Open Library of Humanities

Linking gestural representations to syllable count judgments: A cross-language test

Anisia Popescu, Department of Linguistics, University of Potsdam, Potsdam, Germany; Department of Linguistics, University of Southern California, Los Angeles, CA, USA, anisia.popescu@universite-paris-saclay.fr

Ioana Chitoran, Department of Linguistics, Université Paris Cité, Paris, France; CLILLAC-ARP, France, ioana.chitoran@u-paris.fr

A special class of English words with tense vowel/diphthong nuclei and liquid codas receive variable syllable count judgments (one and over-one syllables). Tilsen and Cohn (2016) showed that differences in judgments correlate with differences in production, supporting their hypothesis that meta-phonological judgments and speech motor control share a common representation. In the present study, we further propose that syllable count judgments are related to subsegmental representation in the rime, and are independent of acoustic duration. We test the hypothesis by comparing English and German, chosen for their similar word structures and vowel length contrast, and their crucial difference in the gestural specification of coda liquids. In English, coda liquids have an earlier vocalic gesture relative to the consonantal one, while in German, both gestures are simultaneous. We stipulated that sesquisyllabic (over-one) judgments are related to the count of sequentially-timed vocalic gestures in the rime. The difference in the coda liquid composition between the two languages predicts that sesquisyllables should not emerge in German. Our predictions were confirmed by the results of parallel production (acoustic) and syllable count judgment experiments in English and German. We propose a model accounting for these results, and we discuss its typological implications and its limitations.

Laboratory Phonology: Journal of the Association for Laboratory Phonology is a peer-reviewed open access journal published by the Open Library of Humanities. © 2022 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/4.0/.
OPEN ACCESS

1. Introduction

Speakers have consistent intuitions about the number of syllables in a word. Even children as young as 4–5 years old can agree on the number of syllables they count in syllable clapping tasks (English: Ziegler & Goswami, 2005; Burt, Holm & Dodd, 1999; German: Fricke, Szczerbinski, Fox-Boyer & Stackhouse, 2016). There is, however, a specific subset of English words that has received much attention in the literature for prompting variable syllable count judgments (Lavoie & Cohn, 1999; Cohn, 2003; Tilsen & Cohn, 2016). These are words whose rimes contain a tense vowel or a diphthong followed by a post-vocalic liquid, such as *feel, file, fear, fire*. For such words, American-English speakers disagree in their syllable count judgments, and attribute *1 syllable, 2 syllables,* or, if given the option, *1.5 syllables* to this particular subset of words. Lavoie and Cohn (1999), who first identified this subclass of words, termed them 'sesquisyllables,' emphasizing the one-and-a-half syllable metric. In their original study, the variability observed in syllable counts is related to acoustic duration: Liquid codas are shown to contribute more to syllable duration than stop consonants do. This finding is interpreted as supporting a trimoraic representation of sesquisyllables (Cohn, 2003; Cohn & Lavoie, 2003). The argument builds on the minimal word requirement present in many languages. In English, as in other languages, content words must be minimally bimoraic (Hayes, 1989; Hammond, 1999). Cohn (2003) argues that, in sesquisyllables, the post-vocalic liquid accounts for an extra weight unit, making them superheavy, thus trimoraic.

In a subsequent study, Tilsen and Cohn (2016) further examined sesquisyllables and found that rime durations and formant trajectories differed for such words depending on whether they were judged as monosyllabic or disyllabic. The authors interpreted these results as evidence for a shared representation between speech motor control and meta-phonological judgments. If Tilsen and Cohn are correct, the correlation between speech production and syllable-count judgments should hold cross-linguistically.

Our goal in the present study is twofold: to test Tilsen and Cohn's interpretation by submitting it to a cross-language comparison, and to propose a gestural account of sesquisyllables cross-linguistically. More specifically, we propose that language-specific differences in speaker intuitions about syllable count judgments are related to the language-specific differences in the gestural organization of coda liquids. To test this hypothesis, we compare two languages with similar phonological word structure, but different gestural specifications of their coda liquids: American English and German. Both languages have CVC words contrasting in vowel length (tense versus lax), thus allowing a direct, one-parameter difference comparison: English has dark [ɫ] in coda position (Sproat & Fujimura, 1993), while German has clear coda [l] (Geumann, Kroos & Tillmann, 1999).

An earlier study limited to syllable count judgments by British English and German native speakers (Popescu, 2019) confirmed a language-specific difference in native speaker intuitions.

British English native speakers, just as American English native speakers, differentiate between sesquisyllables with liquid codas and their structural counterparts involving nasal and stop codas. However, no difference based on coda type was observed for German. Popescu (2019) hypothesized that this distinct intuitive behavior may be attributed to differences in the articulatory composition and the gestural timing of coda liquids between English and German. In the present study, we test this hypothesis by comparing (acoustic) production data, in addition to syllable count judgment data from a single group of native speakers of American English and of German, respectively.

We begin by presenting the gestural specifications of coda liquids in the next section, Section 1.1. The hypotheses and predictions for English and for German are presented in Section 1.2. The remainder of the paper is structured as follows: Section 2 describes the experimental method, Sections 3 and 4 present the results of the English and the German experiments, respectively, and the last two sections focus on the model proposed to account for them (Section 5) and the discussion of its implications (Section 6).

1.1. Gestural composition of coda liquids

Traditionally two varieties of laterals have been described across languages: clear and dark /l/. The two varieties mainly differ in their tongue body configurations and gestural coordination patterns. American English has both varieties, depending on syllable position: clear /l/ in onset and dark /l/ in coda position. German has clear /l/ in both syllable positions. This difference in coda position is crucial to the present study. Clear and dark /l/ have been extensively described for American English (Sproat & Fujimura, 1993; Narayanan, Alwan & Haker, 1997; Hardcastle & Barry, 1989; Browman & Goldstein, 1995; Proctor et al., 2019). The dark/velarized [ɫ] is produced with an earlier occurring tongue dorsum (TD) retraction gesture towards the uvular region. The clear [l] in onset position involves a tongue tip (TT) gesture which is either synchronous to, or slightly precedes a TD lowering (not retraction) gesture. To our knowledge, there are no comparable studies investigating the specific articulation of German clear [l]. Whether German clear [l] has a double gesture or just one (i.e., no specific TD gesture) is still under debate. However, based on the articulatory characteristics of the American English light [l], it is generally assumed that the German lateral consists of synchronous TD and TT gestures (Recasens, 2012; Recasens Pallars & Fontdevila, 1998; Geumann et al., 1999; Marin & Pouplier, 2010; 2014). The present study relies on acoustic data, thus will not contribute directly to this debate. We assume for now, crucially for our analysis, that the TD gesture of the German clear [l], if present, does not precede the TT gesture, as in the case of American English dark [ɫ].

The two types of laterals can be compared acoustically. The main acoustic correlate of lateral darkness is F2, associated with the half-wavelength resonance of the back cavity. In dark [ɫ], F2 values are low, corresponding to either a lowered predorsum, a TD retraction, or both—all

configurations resulting in the lengthening of the back cavity (Recasens, 2012; Narayanan et al., 1997; Stevens, 1998; Fant, 1960). In clear [l], F2 is higher corresponding to the absence of TD retraction, and thus no increase of the back cavity.

Figure 1 illustrates the difference between English and German coda laterals in three pairs of words from our data, sharing the same vocalic nuclei: English feel [fiːɫ], pool [puːɫ], tile [taɪɫ], and German viel [fiːl], Stuhl [ʃtuːl], Teil [taɪl]. The difference in lateral quality between the two languages is visible in the formant trajectories during the vowel-lateral (VL) rime. In all three examples in **Figure 1**, formant values start in similar ranges at the beginning of the vowels, but trajectories subsequently shift to low F2 for English, indicating tongue predorsum lowering, tongue dorsum retraction, or both, and higher F2 for German. Even though exact tongue configurations cannot be derived, the formant trajectories seen in **Figure 1** are a clear indication that the tongue dorsum component in our English and German laterals is quite different.

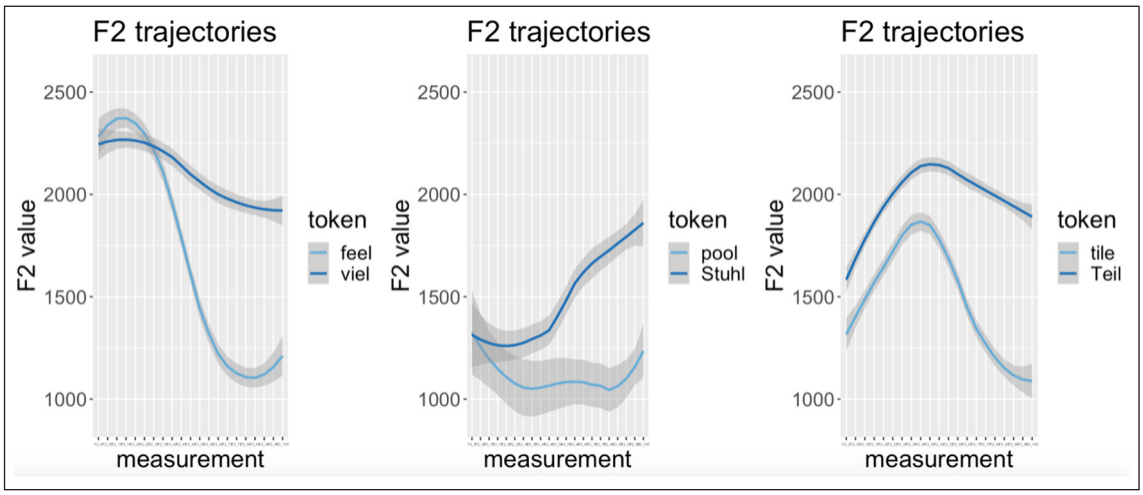

Figure 1: Smoothed F2 trajectories in similar English (light blue) and German (dark blue) words: feel [fiːɫ] versus viel [fiːl] 'much' (left); pool [puːɫ] versus Stuhl [ʃtuːl] 'chair' (center); tile [taɪɫ] versus Teil [taɪl] 'part' (right). Formant values extracted at 20 equally spaced steps across the vowel-lateral sequence in the acoustic data recorded for the present study.

The dark and clear laterals also differ in coarticulation degree. Proctor (2009), Proctor and Walker (2012) found that dark [ɫ] is more resistant to coarticulation than clear [l], indicating that clear [l] is more permissive in its articulatory configuration. This is consistent with the prediction of the degree of articulatory constraint (DAC) model of Recasens, Pallars, and Fontdevila (1997), where coarticulation resistance increases with the degree of involvement of the tongue dorsum.

Since for clear [l] the tongue dorsum gesture plays only a minor role (Recasens et al., 1998) one could argue that the tongue dorsum is not actively involved in the production of clear [l], as it is in the production of dark [ɫ]. If there is no specific target for the tongue dorsum itself, it remains more available for coarticulation. More recently, however, evidence of active control of the tongue blade during the production of lateral channels during the production of the lateral has been shown for Australian English (Ying et al., 2021). Also, specifically for German, Pastätter and Pouplier (2014) showed that specifying a tongue body target for all coronal consonants, including the clear lateral, is necessary for the accurate modeling of German coronal consonants using the TADA task dynamic synthesizer (Nam, Goldstein, Saltzman & Byrd, 2004), which is the computational implementation of the Task Dynamics model (Saltzman & Munhall, 1989). This result, however, should be interpreted with caution, as it is highly dependent on the specifications of TADA.

The liquids we are concerned with also include the rhotics. The American English rhotic is an approximant, and, like the lateral, is a complex segment composed of a double lingual gesture (a tongue tip/body and a tongue root gesture) and a labial gesture (Mielke et al., 2016; Proctor et al., 2019; Campbell, Gick, Wilson & Vatikiotis-Bateson, 2010; Gick & Campbell, 2003; Gick & Goldstein, 2002; Alwan, Narayanan & Haker, 1997; Zawadzki et al., 1980; Delattre et al., 1968). Two main tongue configurations are attested for American-English rhotics: retroflex, which involves a raised and curled-up tongue tip with a lowered tongue dorsum, and bunched, with a lowered tongue tip and raising of the tongue body. Proctor (2009) showed that, in both varieties, the tongue root gesture serves as a stabilization anchor for the tongue tip gesture. The tongue root gesture occurs closer to the vocalic nucleus, following the labial and tongue tip gestures in onsets and preceding them in coda position (Campbell et al., 2010; Gick & Goldstein, 2002). This implies that in rhotics, as in the case of laterals, the tongue root gesture occurs either synchronously or before the tongue tip gesture in coda position. In German, the rhotic shows high dialectal variability (Göschel, 1971; Wiese, 2003). There are two main dialectal variants of the German rhotic—an apical and a uvular trill—which can be produced as taps, fricatives, approximants, or vowels as a consequence of articulatory reduction (Mooshammer & Schiller, 1996; Schiller & Mooshammer, 1995; Hall, 1993). The German rhotic—in either onset or coda—is dialect- and speaker-dependent.

Finally, and relevant for our hypothesis, the tongue dorsum gestures involved in the production of the English liquids are considered to be vocalic, similar to vowel gestures (Giles & Moll, 1975; Sproat & Fujimura, 1993; Gick, Min Kang, & Whalen, 2002). Independently, similar intuitions about a partly vocalic quality of the liquids are captured in several formal accounts of syllable structure. Harris (1994) and Green (2001), for example, have proposed that coda

liquids are better analyzed as forming diphthongs with the nucleus. Psycholinguistic evidence from word games (Treiman, 1984) and from speech errors (Stemberger, 1981) also support a representation of post-vocalic liquids as part of a complex vowel, rather than patterning with coda consonants.

To summarize, the articulation of coda liquid consonants differs between English and German. English coda laterals and rhotics are produced with two sequential gestures, the vocalic tongue dorsum gesture preceding the consonantal tongue tip gesture. The German coda lateral is possibly also produced with a double gesture, but the tongue dorsum gesture is less actively involved, and does not precede the tongue tip gesture. The hypotheses and predictions presented next crucially take into account the gestural composition of the liquids.

1.2. Hypotheses and predictions

Assuming gestural representations, we claim that the *relative timing* and the *quality* of the gestures (vocalic versus consonantal) involved in the production of coda liquids play a role in speakers' intuitions about syllable counts. More specifically, the earlier timing of the liquid vocalic gesture in the coda, adjacent to the vowel gestures of the syllable nucleus, can account for the non-standard over-one syllable count judgments given to sesquisyllables in English. In other words, a syllable containing a tense vowel or diphthong and a coda liquid would consist of three vocalic gestures that are sequentially coordinated. In moraic terms, these would form a 'superheavy' structure which would interfere with the standard prosodic weight of monosyllabic words (heavy syllable). This would mean that only vocalic gestures would contribute to syllable weight, accounting for the Weight by Position rule. In German, however, where laterals have no sequentially timed vocalic gesture, a superheavy nucleus would not be formed. The absence of such a nucleus translates into the absence of sesquisyllables in German.

Our main theoretical hypothesis is that speakers' intuitions about syllable counts are linked to the gestural composition of coda liquids (in languages with vowel length distinction). Thus, we postulate that sesquisyllables involving liquid consonants are only present in English and not in German. Furthermore, we postulate that, in a syllable count task, even when participants attribute just one syllable to such words, reaction times should reflect the presence of sesquisyllables in English but not in German. For tokens involving liquid codas, English speakers should take longer to decide, resulting in longer reaction times than for non-liquid codas. Finally, we hypothesize that duration is not the only factor influencing speakers' syllable count judgments (henceforth 'SCJ'). Our theoretical and experimental hypotheses are summarized in **Table 1**.

Main hypothesis	
Speakers' intuitions about syllable counts are linked to coda liquid gestural composition	
SCJ	
Experimental hypothesis	Alternative hypothesis
Rimes with coda laterals (and rhotics) count as sesquisyllables in English	Rimes with coda laterals count as sesquisyllables in both English and German
Reaction times	
Experimental hypothesis	Alternative hypothesis
Reaction times are longer for rimes involving liquid codas in English	Reaction times do not differ based on rime composition in either English or German
Duration	
Experimental hypothesis	Alternative hypothesis
Duration does not correlate with SCJ	Duration correlates with SCJ

Table 1: Main SCJ and duration hypotheses.

Based on our hypotheses, we make the following predictions about syllable count judgments, rime duration, and their correlation.

First, the predictions pertaining to SCJs are as follows. For English, we expect over-one SCJs to be attributed predominantly for liquid codas. Furthermore, based on Tilsen and Cohn's (2016) findings, we expect syllables involving diphthongs to receive more ambiguous SCJs than their tense vowel counterparts. In the case of German, we predict no distinguishable SCJ pattern based on coda type. We thus expect that similar percentages of SCJs (whether one syllable or more than one syllable) are attributed to all coda types.

Second, rime duration predictions can be made in relation to our SCJ prediction. If our hypothesis is correct, and sesquisyllables can be predicted by the gestural composition of coda consonants, we do not expect duration to be correlated with SCJs over all rime classes. Rather, in English, we expect rime duration to correlate with SCJs only within the subset of words with coda liquids: Words that receive over-one syllable count judgments among this type of words should exhibit longer rimes than words judged to be monosyllabic.

The cross-language comparison will be presented as two independent experiments. Experiment 1 is an extension of Tilsen and Cohn's (2016) study of American-English, including a larger variety of stimuli. Experiment 2 presents new data from German native speakers. The experimental protocol, the acoustic analysis, and the statistical analysis are identical for both experiments and are presented in the following section.

2. Methods

Twenty American-English native speakers, undergraduate students at the University of Chicago, and 16 German native speakers, students at the University of Potsdam, participated in the study. The American participants were recruited as volunteers in Chicago; the German participants received either payment or course credit in Potsdam. None of the participants reported any history of hearing or language impairment. All participants gave written informed consent for their participation in the experiment and for the subsequent use of their data for scientific purposes.

2.1. Experiment design

The same experiment design was used in both locations, in the US and in Germany, respectively. Following the experimental protocol of Tilsen and Cohn (2016), the study was organized in two sequential tasks: a speech production task, immediately followed by a syllable count judgment task. Crucially, participants were unaware of the upcoming judgment task at the time of the production experiment.

For the recordings, participants were asked to read target, control, and filler words embedded in carrier phrases: (English) "I say … now"/ (German) "Ich sage …. drei mal." Stimuli, described in the next section, were identical across tasks.

2.1.1. Stimuli

The English stimuli were based on Tilsen and Cohn (2016), with the following additions: Lateral coda targets included two additional nuclei (/uː/ and /eɪ/), and we included more variation in coda stop controls (/k/, /d/, /dʒ/) and coda nasal controls (/m/, /n/). **Table 2** contains the complete list of target (T) and control (C) stimuli. The rime combinations include [iː, uː, aɪ, eɪ] syllable nuclei and one of five possible coda types: no coda (open), stop and nasal coda for controls, and lateral and rhotic coda for targets.

As far as the lexicon allowed it, the orthography was controlled for: Most stimuli with a tense vowel + C are spelled CVVC, most stimuli with a diphthong + C are spelled CVCV, and all those with nasal codas are spelled CVVC. Only four stimuli—*gleam, claim, stay, spook*—have complex onsets. Control for onset complexity was preferred given our hypothesis, because in English, syllables with complex onsets are known to have shorter vowel durations compared to their simplex onset counterparts (Browman & Goldstein, 1988; Honorof & Browman, 1995; Byrd, 1995).

nucleus	coda				
	open (C)	**stop (C)**	**nasal (C)**	**lateral (T)**	**rhotic (T)**
iː	bee fee pea	feed peak	keen lean beam gleam team	feel heel wheel	beer pier fear gear
uː	Pooh zoo	spook	zoom	pool	
aɪ	tie			vile tile	fire tire liar wire
eɪ	may stay	page	claim pain	male pale	
Total	8	4	8	8	8

Table 2: English target (T) and control (C) stimuli per nucleus and coda type.

The German stimuli were similar in structure to the ones presented in Experiment 1. Nuclei were more varied, including all six long monophthongs of German (/aː, eː, iː, oː, uː, yː/) and two diphthongs /aɪ/ and /aʊ/. All diphthong and most tense vowel tokens are spelled CVVC; a few tokens are spelled CVhC. As for English, complex onsets were avoided as much as possible, because vowel compression effects have been found for German complex onsets, as well (Marin & Pouplier, 2010; Pouplier, 2012). In our list, only five words contain the complex onset [ʃt]. **Table 3** shows all targets (T) and controls (C).

nucleus	coda				
	open (C)	**stop (C)**	**nasal (C)**	**lateral (T)**	**rhotic (C)**
aː	sah [zaː] ‘see, 3sg-pst’	Staat [ʃtaːt] ‘state’	Wahn [vaːn] ‘delusion’	Saal [zaːl] ‘room’	Haar [haːr] ‘hair’
aʊ	Sau [zaʊ] ‘female pig’		Baum [baʊm] ‘tree’	faul [faʊl] ‘lazy’	

(Contd.)

nucleus	coda				
	open (C)	stop (C)	nasal (C)	lateral (T)	rhotic (C)
			Raum [ʁaʊm] ‘room’ Faun [faʊn] ‘faun’		
aɪ	sei [zaɪ] ‘be’	Teig [taɪk] ‘dough’ seit [zaɪt] ‘since’	Heim [haɪm] ‘home’ sein [zaɪn] ‘to be’ mein [maɪn] ‘my’	Teil [taɪl] ‘part’ Seil [zaɪl] ‘rope’ Heil [haɪl] ‘well being’ feil [faɪl] ‘venal’	
eː	Fee [feː] ‘fairy’ See [zeː] ‘sea’		Fehn [feːn] ‘mire’	fehl [feːl] ‘amiss’ Mehl meːl] ‘flour’	sehr [zeːʁ] ‘a lot’
iː	die [diː] ‘the,fem’ Vieh [viː] ‘cattle’	Lied [liːd] ‘song’ stieg [ʃtiːk] ‘climb, 3sg-pst’	Wien [viːn] ‘Vienna’	viel [fiːl] ‘a lot’ Stiel [ʃtiːl] ‘flower stalk’ Kiel [kiːl] ‘keel’ Ziel [tsiːl] ‘goal’	vier [fiːʁ] ‘four’ Tier [tiːʁ] ‘animal’ Stier [ʃtiːʁ] ‘bull’ Bier [biːʁ] ‘beer’
oː				wohl [voːl] ‘well’	Chor [koːʁ] ‘choir’

(Contd.)

nucleus	coda				
	open (C)	stop (C)	nasal (C)	lateral (T)	rhotic (C)
uː				Stuhl [ʃtuːl] 'chair'	fuhr [fuːʁ] 'drive, 3sg past'
yː		blüht [blyːt] 'bloom, 3sg-present'		kühl [kyːl] 'cool'	
Total	6	6	9	15	8

Table 3: German target and control stimuli per nucleus and coda type.

For both English and German, fillers consisted of unambiguously monosyllabic and disyllabic words. The full list of fillers is given in Appendix A.

2.1.2. Production task

Two repetitions of each phrase were elicited in two randomized blocks. Participants were asked to read the sentences clearly, at their normal speech rate, without over-emphasizing the variable word in the carrier phrase. The productions were monitored by the experimenter, and participants were asked to repeat any sentences which didn't follow the instructions or seemed problematic. Recordings took place in the soundproof booths of the Linguistic Departments of the Universities of Chicago and Potsdam, respectively. The same Zoom H4NPRO recorder was used for all recordings.

2.1.3. Syllable count judgment task

For the SCJ task, participants were asked to decide how many syllables there are in a word by choosing one of three answer options: 1, 1.5, or 2 syllables. Each word was presented in written form on a computer screen, and SCJ responses were recorded through a C# application, created specifically for this task to manage the pseudo-randomization, the timing of stimuli presentation, and the recording of the response time per answer. The same application was used to record speaker information such as age, gender, and language background.

While the instructions for the SCJ task were the same as in Tilsen and Cohn (2016), some experimental details were modified as follows. SCJ elicitation was presented as discrete answer options, as opposed to a continuous scale. This modification was recommended by Sam Tilsen (personal conversation) in order to avoid the difficulties of mapping gradient to categorical SCJs encountered in their experiment. Contrary to Tilsen and Cohn (2016), who discarded all answers which took more than 5 seconds, no time limit was imposed for the SCJ task, reaction times being recorded instead.

Each word appeared on the screen for 1.5 seconds, enough time for participants to read the word silently and to subvocalize it. Participants were specifically asked to subvocalize each word and think about how they would produce it before giving their answer. This step in the task was added so that speakers would rely less on the orthographic representations of the words, and more on their own proprioceptive articulatory and auditory feedback. The word disappeared after 1.5 seconds, and three buttons, one for each answer option, appeared on a horizontal line. At the start of each trial, the mouse cursor appeared in the middle of the screen, below the 1.5 answer option, at equal distance from the 1 and 2 answer options, as shown in **Figure 2**. Once participants clicked on one of the three buttons, a "Next" button appeared, which prepared them for their next trial. Reaction times were recorded from the moment the three answer options appeared on the screen to the moment the participants clicked on one of the response options. Trials were pseudo-randomized in order to avoid making a judgment by directly comparing tense vowel to diphthong nuclei words. The experimental block was therefore structured as follows: Trials with tense vowel nuclei (targets and controls), mixed with part of the fillers, were presented first. Once all tense vowel nuclei trials were exhausted, trials with diphthong nuclei (targets and controls) were presented together with the remaining fillers.

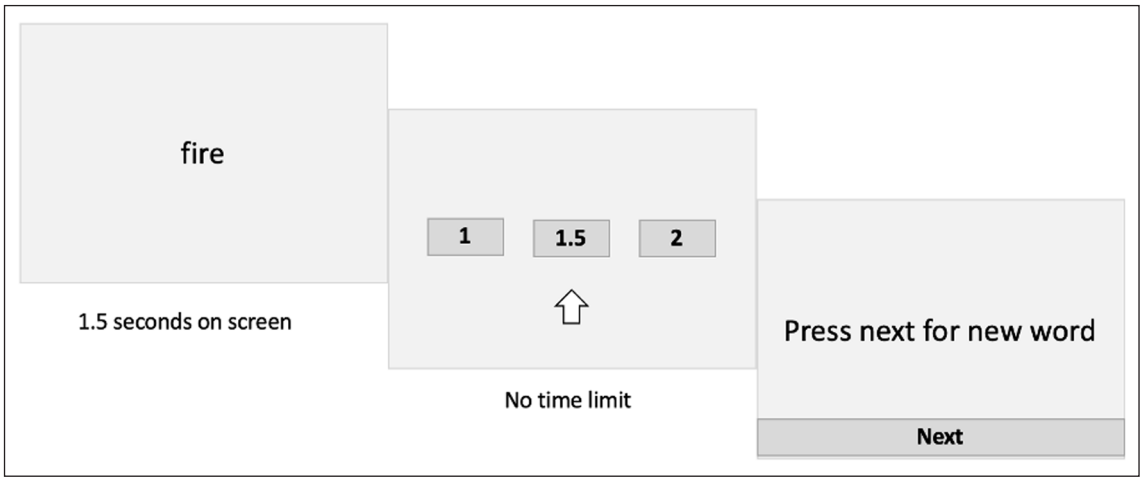

Figure 2: Schematic representation of the experimental progress of one trial of the SCJ task.

In order to justify the option of 1.5 syllables, our written instructions followed Tilsen and Cohn (2016), specifying that: 'There is no right answer. People may disagree on syllable counts for certain words, and sometimes people feel that the number of syllables in a word is between whole numbers." The written instructions were further clarified orally by the experimenter, who explained that the 1.5 syllable option is a stand-in value for words considered to have a syllable count in-between one and two syllables.

Participants who gave more than 15% non-standard SCJs for unambiguous disyllabic words were excluded. This includes, for example, participants who chose options 1 or 1.5 syllables for a word such as *doctor*. This was the case for two of the 20 American English native speakers, thus leaving a total of 18 participants whose data were analyzed.

2.2. Acoustic data analysis and statistical analyses for the production task

For the production study, the response variable considered is acoustic rime duration, corresponding to vowel duration in open syllables, and to vowel-consonant duration in closed syllables. All acoustic files were hand segmented in Praat (Boersma & Weenink, 2019). Syllable rimes were labeled based on the waveform and wide-band spectrogram. The vowel onset was marked at the onset of F1. In open syllables, the vowel F2 offset was taken as the end of the rime. For rimes containing post-vocalic liquids, the end of visible formant structure on the spectrogram marked the end of the rime. Most post-vocalic word-final stops were released, and the end of the rime was marked at the offset of the release burst. When stops were not released, the onset of formant structure for the nasal of 'now' was marked as the final boundary. For word-final fricatives and affricates, the change in waveform profile at the offset of post-vocalic frication was taken as final boundary. **Figure 3** gives examples of the segmentation for the words *pale, liar, feed,* and *page*, each from a different speaker.

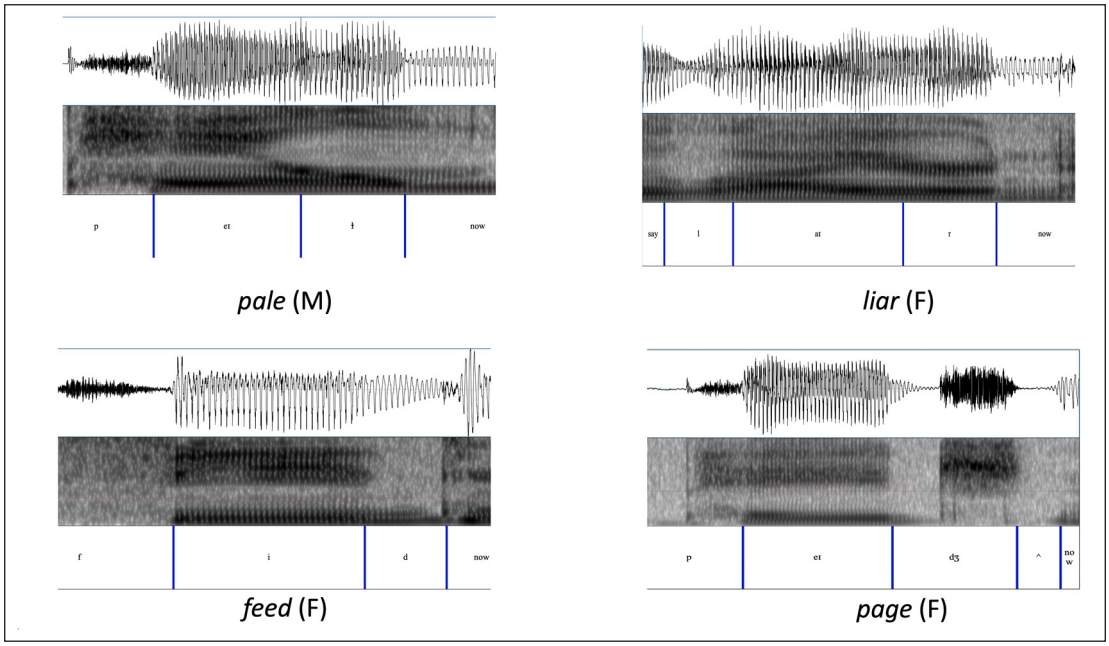

Figure 3: Labeled waveforms and spectrograms for the words *pale, liar, feed, page* for four different speakers (M – male, F – female).

In the case of postvocalic nasals, to separate a word-final [n] from the onset [n] of the following word 'now,' we relied either on changes in waveform and/or spectrogram, or on the short pause between the two nasals if it was present. **Figure 4** shows examples of the nasal segmentation for the words *keen, lean,* and *beam*. The words *keen* and *beam* are examples of a segmentation based on visible changes in the waveform and/or spectrogram. The word *lean* shows a case where the participant pauses between the target word and the following word *now*. The segmentation was done by the first author, double checked by the second author.

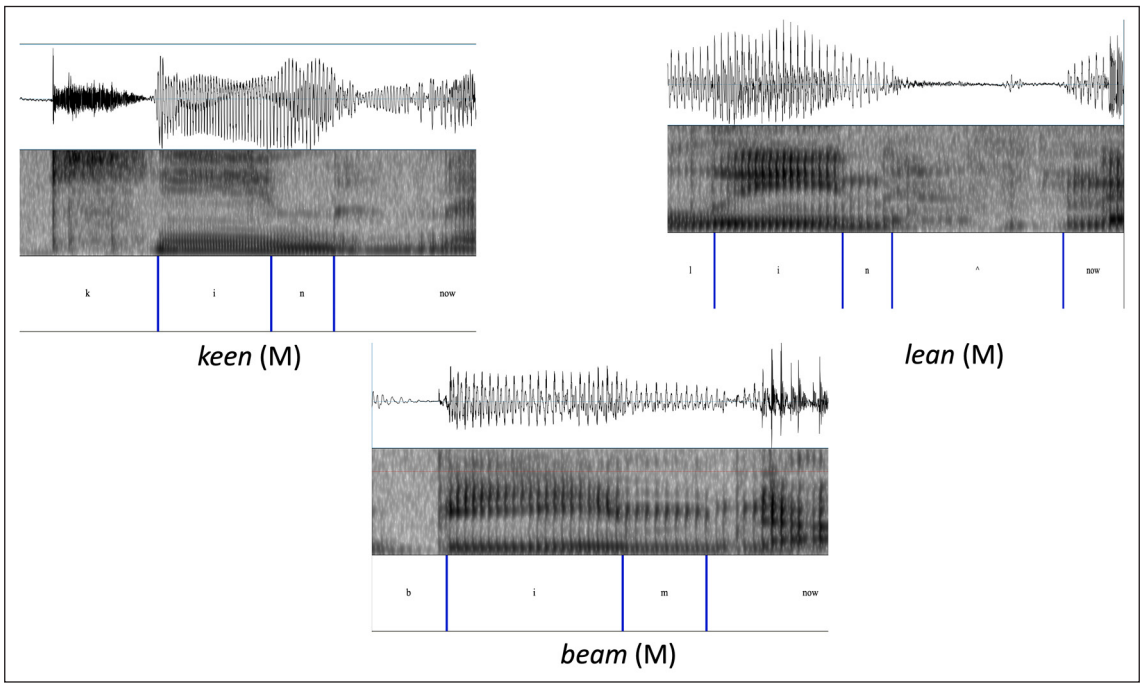

Figure 4: Labeled waveforms and spectrograms for the words *keen, lean,* and *beam* for four different speakers. (M – male, F – female).

Two different metrics for rime duration were considered. The first corresponds to the relativized rime duration measure used in Tilsen and Cohn (2016). We followed their normalization procedure, and we divided the duration of a rime (for each speaker and repetition) by the mean duration of the corresponding open syllable rime. For example, the mean duration of /i/ in open syllables (*fee, pea, bee*) was used as a denominator of the ratio for the duration of /il/ in *feel*, /im/ in *beam*, and /id/ in *feed*. For the German data, since the wordlist did not always include open-closed syllable rime pairs (there are no open syllables including /y:, u:, o:/), we divided the rime duration in closed syllables by the rime duration in open syllables containing same height vowels (i.e., /y:C/ and /u:C/ divided by /i:/, and /o:C/ divided by /e:/). As stated by Tilsen and Cohn (2016), this measure is meant to take into account differences in speakers' baseline word duration and speaking rate when measuring the contribution of the coda consonant to the rime.

The second measure we considered was the average of raw rime duration over two repetitions for each speaker. This measure does not control for the additional factors mentioned, but is useful in allowing for a more direct interpretation of the numerical values in ms rather than ratios, while averaging over tokens.

Rime duration data was analyzed using linear mixed effects models (*lme4* package, Bates, Mächler, Bolker, & Walker, 2015). Models were run separately for each of the two metrics used. Individual rimes (rime_identity) were used as a main predictor. A second predictor was lexical frequency, extracted from the Corpus of Contemporary American English – COCA (Davies, 2008) and from a 100,000-word German Wikipedia corpus downloaded from the Wortschatz Universität Leipzig site (Goldhahn, Eckart & Quasthoff, 2012). Random factors were Participant for the averaged rime duration models, and Participant and Repetition for the normalized rime duration models. All models included random intercepts. Pair-wise comparisons between levels were conducted by releveling factors.

The correlation between syllable count judgments and duration was analyzed using cumulative link mixed effects models (*ordinal* package: Christensen, 2019) with rime_duration and coda_type as fixed factors and Participant and Repetition as random factors. Random intercepts were included in all models.

For the reaction time data, a whole distribution analysis of reaction times was preferred (Whelan, 2008) using *glmer* (*lme4* package Bates et al., 2015) with gamma distribution (which best fit our distributions) and log link functions. Mean, median, and standard deviation measures are also provided. Outliers were not eliminated, consistent with the task instructions, namely the fact that participants were explicitly asked to subvocalize and think about the pronunciation of each word before making their choice.

For all models, the significance of the main effects was tested using chi-square likelihood ratio tests. Model diagnostics plots for the final models were analyzed to test for deviations from homoscedasticity or normality. The *lmerTest* library (Kuznetsova et al., 2017) was used to determine significance levels.

3. Experiment 1: American English

We first present the SCJ measure and response times for SCJs per coda type in Section 3.1. Rime duration measures are presented separately, in Section 3.2. For both SCJ and duration measures, an individual rime analysis will be presented. The last section (Section 3.3) presents the correlation between SCJs and duration.

3.1. Results: Syllable count judgments

As predicted, ambiguous over-one SCJs were attributed exclusively to liquid coda rimes. Open syllables as well as closed syllables with post-vocalic nasals or stops received only

monosyllabic judgments. Furthermore, and consistent with Tilsen and Cohn's (2016) results, target words involving diphthongs received more over-one SCJs than those involving tense vowels.

Three of the 18 participants consistently attributed over-one SCJs to all rimes with liquid codas (*wheel* being the only exception). Three more participants awarded over-one SCJs to specific targets, mainly involving diphthong nuclei. All of the remaining participants (n = 12) awarded over-one SCJs at least once, to diphthongs followed by a liquid coda (most frequently *liar* and *fire*). **Figure 5** shows the counts of attributed SCJ options per coda and nucleus type. **Figure 6** shows a more detailed count of SCJ options per rime identity.

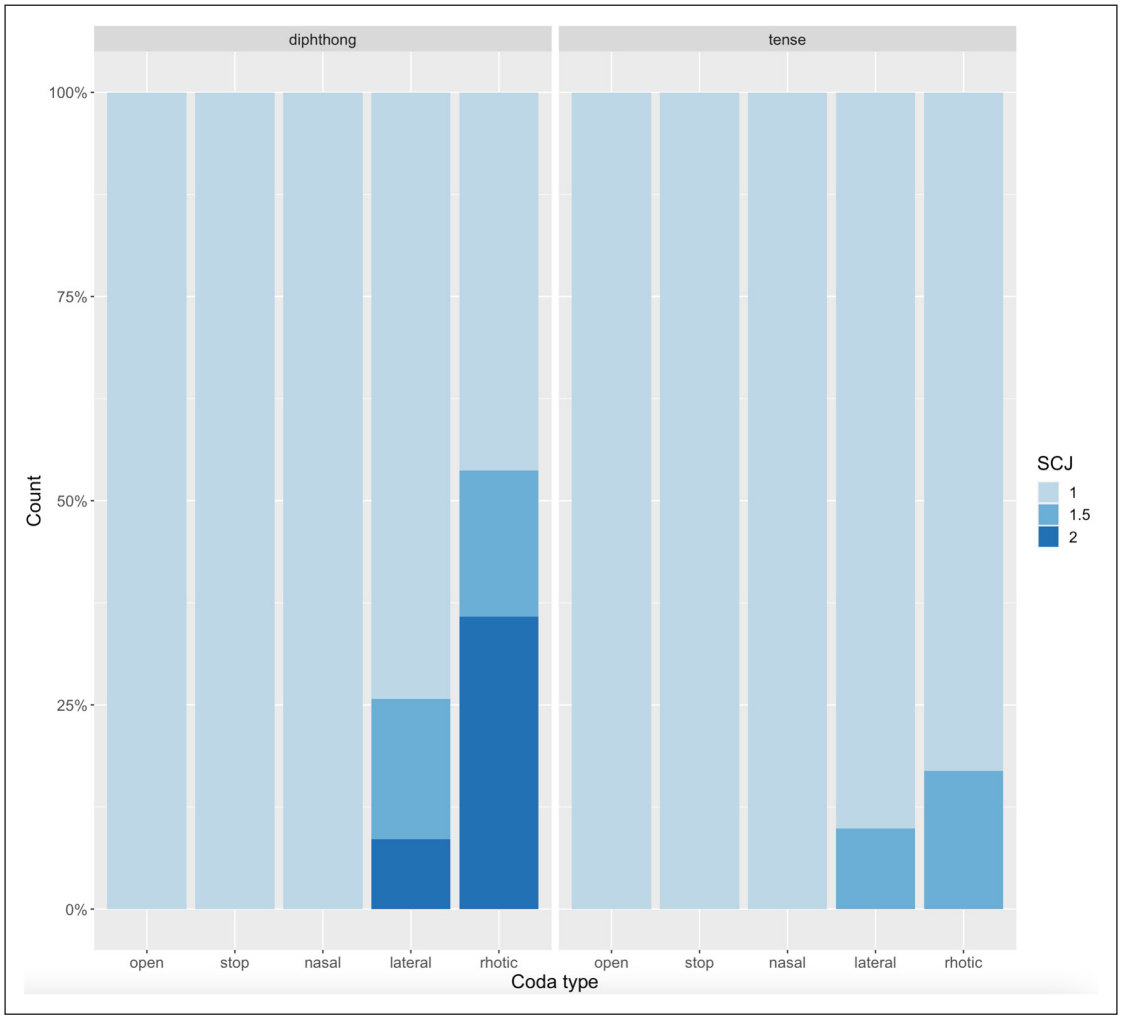

Figure 5: Percentages of attributed SCJs per coda (open, stop, nasal, lateral, rhotic) and nucleus type (diphthong, tense vowel).

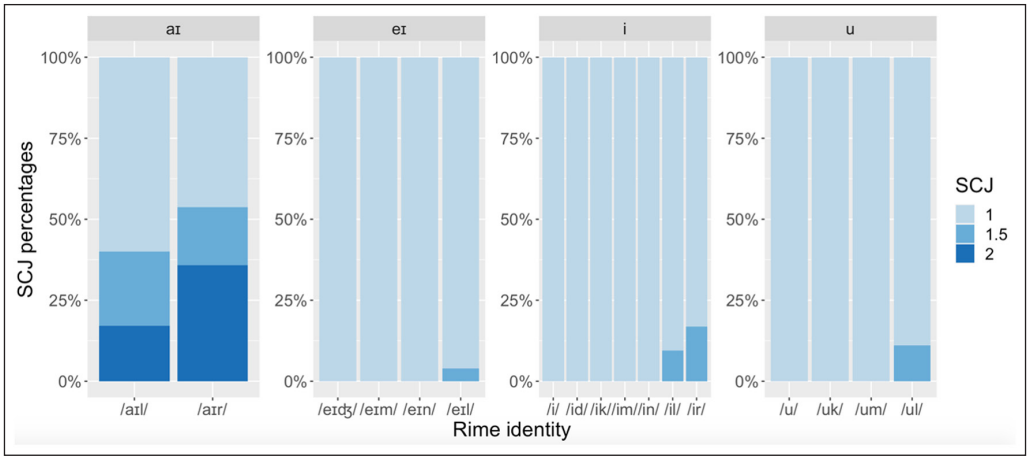

Figure 6: Percentages of attributed SCJs per rime identity and nucleus type (/aɪ, eɪ, i, u/). CVC tokens only.

Table 4 shows the number of participants having attributed over-one SCJs per target token as a function of syllable nucleus and rime identity. The following patterns emerge. First, none of the controls received syllable-counts of over-one. The diphthong target /aɪ/ received the most over-one SCJs, more than tense vowel targets. The 2-syllable answer option was attributed exclusively to tokens involving the diphthong /aɪ/. For tense vowels two of the participants consistently attributed over-one syllable counts. A third participant attributed over-one SCJs to a subset of tense vowel tokens. Rhotic coda targets received more over-one SCJs than their lateral coda counterparts. These patterns confirm Tilsen and Cohn's (2016) findings and extend them to the additional nuclei /u/ and /eɪ/.

Stimuli	**nucleus**	**coda**	**token**	**1.5**	**2**	**Total >1**
Targets	**/i/**	**/il/**	feel	2	0	2
			heal	2	0	2
			wheel	1	0	1
		/ir/	beer	3	0	3
			fear	3	0	3
			gear	2	0	2
			pier	4	0	4
	/u/	**/ul/**	pool	2	0	2
	/aɪ/	**/aɪl/**	vile	3	4	7
			tile	5	1	6

(Contd.)

Stimuli	nucleus	coda	token	1.5	2	Total >1
		/aɪr/	fire	4	4	8
			tire	4	4	8
			wire	3	3	6
			liar	1	11	13
	/eɪ/	/eɪl/	male	2	0	2
			pale	2	0	2
Controls	/i/	Ø	bee	0	0	0
			fee	0	0	0
			pea	0	0	0
		/im/	beam	0	0	0
			gleam	0	0	0
			team	0	0	0
		/in/	lean	0	0	0
			keen	0	0	0
		/id/	feed	0	0	0
		/ik/	peak	0	0	0
	/u/	Ø	Pooh	0	0	0
			zoo	0	0	0
		/uk/	spook	0	0	0
		/u/	zoom	0	0	0
	/aɪ/	Ø	tie	0	0	0
	/eɪ/	Ø	may	0	0	0
			stay	0	0	0
		/eɪdʒ/	page	0	0	0
		/eɪm/	claim	0	0	0
			pain	0	0	0

Table 4: Number of speakers (out of 18) having chosen 1.5 and 2 SCJs per target token, as a function of nucleus type and rime identity.

Response times (RT) ranged from 0.14 to 12.04 seconds, with a mean response time of 1.2 seconds (all coda types included). **Table 5** shows the mean, median, and standard deviation values for each distribution per coda type. **Figure 7** shows the raw distribution of response times as a function of coda type. RT mean, median, and standard deviation values are higher for tokens

involving liquid codas, suggesting that participants take longer to determine syllable counts for sesquisyllables.

	open	**stop**	**nasal**	**lateral**	**rhotic**
Mean	0.96	1.04	1.11	1.42	1.43
Median	0.76	0.83	0.78	1.00	0.96
standard deviation	0.91	0.59	1.41	1.35	1.48

Table 5: Mean, median, and standard deviation for reaction times distributions per coda type.

In addition to the central tendency approach, we ran a whole RT distribution analysis (Whelan, 2008; Luce, 1986). Reaction times faster than 100 ms were eliminated. Very slow reaction times were not eliminated, resulting in right skewed RT distributions for all coda types. We ran a Gamma distribution (which best fit our experimental distributions) with log-link function *glmm* with RT as a response variable and coda type and lexical frequency as predictors. As a random factor we included Participant with random intercepts. Results are shown in **Table 6**.

	Measures	**open**	**stop**	**nasal**
lateral	Estimate	–0.32	–0.27	–0.21
	t-value	–4.80	–2.64	–3.09
	p-value	<0.001	<0.01	<0.01
rhotic	Estimate	–0.39	–0.34	–0.28
	t-value	–6.08	–3.40	–4.21
	p-value	<0.001	<0.001	<0.001

Table 6: Glmm results: Significance levels for RT (log values) as a function of coda type.

The results show no difference in reaction times between laterals and rhotics (Estimate $\sim$ 0.71, *t*-value $\sim$ 1.076, *p*-value $\sim$ 0.28), but tokens involving liquid consonants exhibit higher reaction times than those involving non-liquid consonants, as indicated by the negative estimates in **Table 6**. This suggests that sesquisyllables are processed differently than tokens involving non-liquid codas.

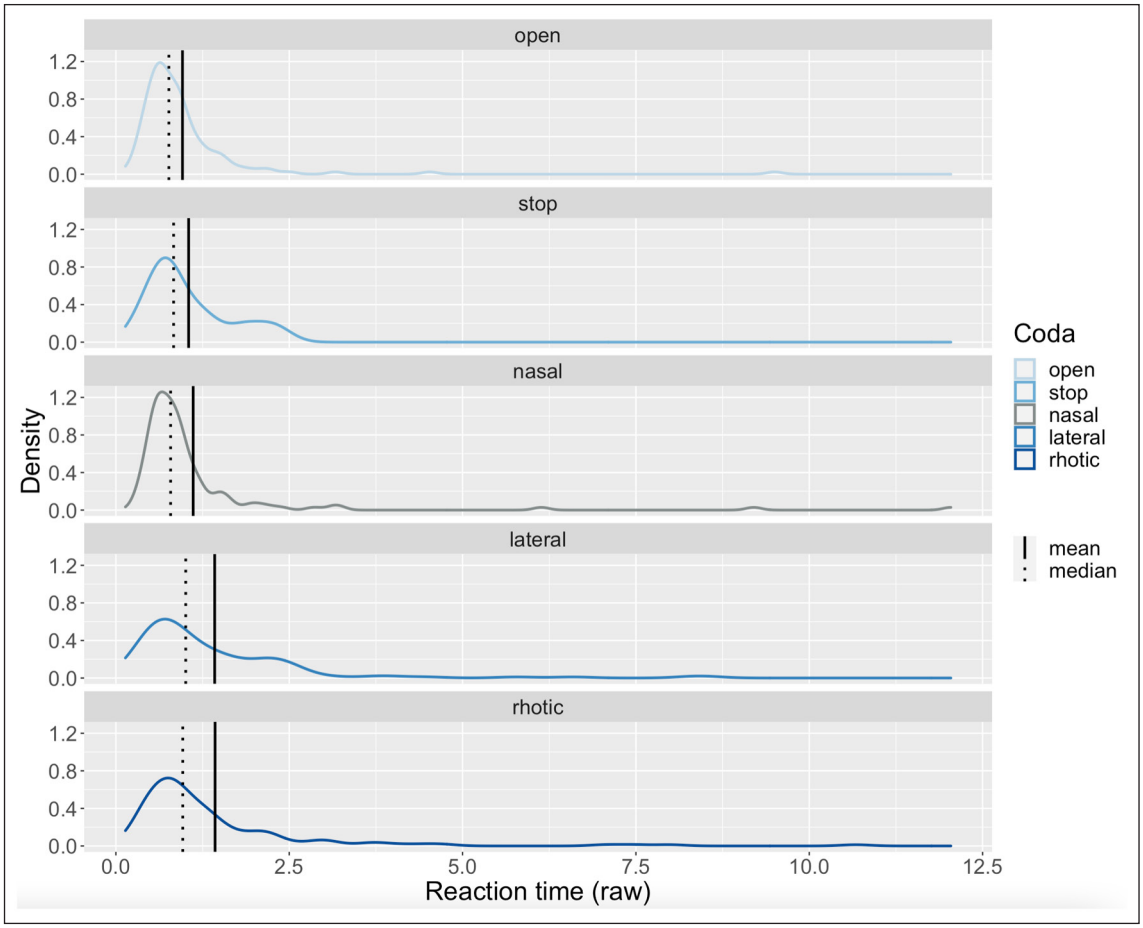

Figure 7: Raw distributions of response times per coda type—outliers included. Vertical solid lines represent the means for each distribution. Vertical dashed lines represent median values for each distribution.

In the next section we present rime duration results.

3.2. Results: Duration

As explained in the Methods section, two measures were considered for the rime duration: The normalized rime duration indicates to what degree the coda consonant contributes to the rime. Higher values indicate higher contributions of the coda consonant. The averaged rime duration, measured in ms as the average over repetitions, gives a directly interpretable metric of rime duration. Recall our prediction: If acoustic rime duration is the determining factor in participants' SCJs, we would expect (1) that targets containing liquid codas would have the longest rime duration overall, and (2) that rimes containing rhotics would be longer than those containing laterals.

Linear mixed models with rime_duration as a response variable and rime_identity and lexical frequency as predictors were considered with each lateral coda token (/aɪl/, /eɪl/, /iːl/, uːl/) as reference levels. Participant was used as random factor for averaged rime duration, and Participant and Repetition were used for the normalized rime duration model. Both models had random intercepts. A likelihood ratio test of the model with rime identity prediction against the null model revealed a significant effect of rime identity (average duration model: $\chi^2(15) = 326.68$, $p < 0.0001$; normalized duration model: $\chi^2(15) = 229.86$, $p < 0.0001$). A likelihood ratio test of the model adding lexical frequency as a predictor showed no significant effect of lexical frequency (average duration model: $\chi^2(1) = 1.97$, $p{\sim}0.16$; normalized duration model: $\chi^2(1) = 2.68$, $p{\sim}0.10$). Consequently, the final models we retained and report on include only *rime_identity* as a predictor. **Figure 8** shows rime duration (both normalized and averaged measures) per rime identity.

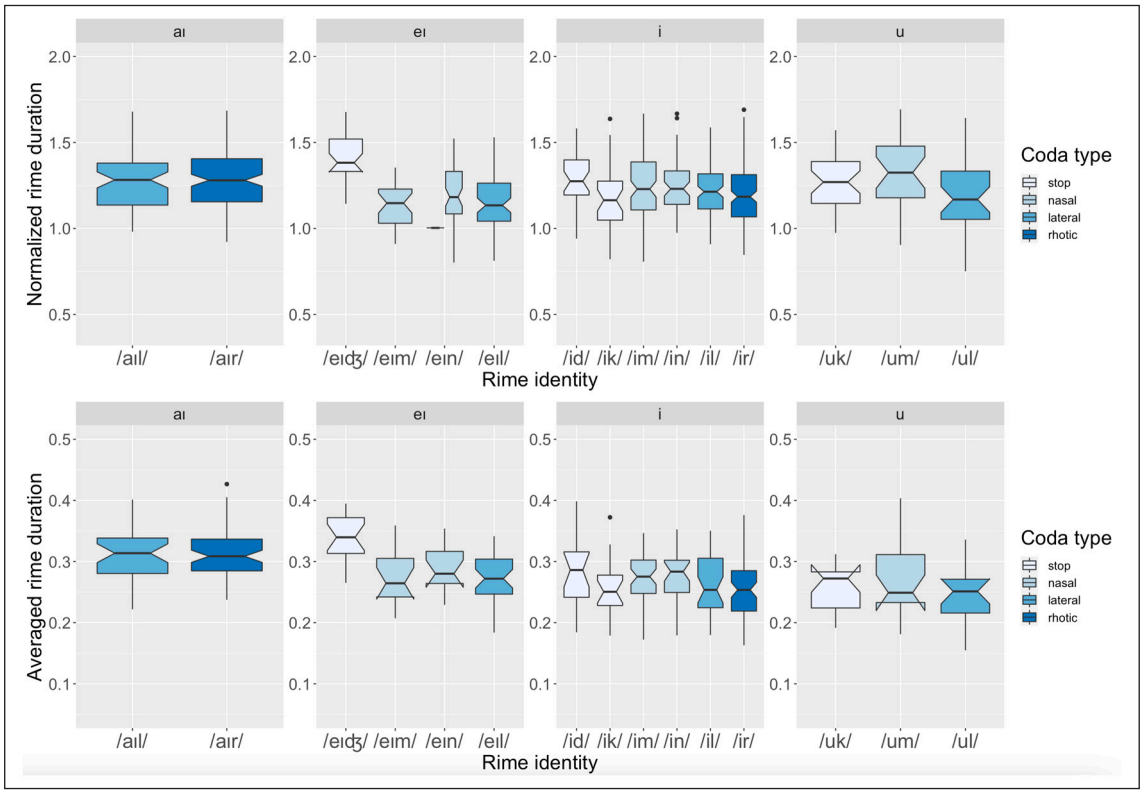

Figure 8: Normalized (upper panel) and averaged (lower panel) rime duration per rime identity, coda type (stop, nasal, lateral, rhotic), and nucleus type (/aɪ, eɪ, iː, uː/). Closed syllable tokens only.

A full list of pair-wise comparisons of rime durations per rime identity are shown in Appendix B. Below, we present crucial results that suggest duration is not the only contributing factor for SCJs:

Words with liquid rimes do not exhibit the longest rime duration. The /eɪdʒ/ rimes are significantly longer than all liquid rimes (/aɪl/: Est ~ –20 ms, *t*-value ~ –3.39, *p*-value <0.001; /eɪl/: Est ~ –69 ms, *t*-value ~ –8.26, *p*-value <0.0001; /il/: Est ~ –78 ms, *t*-value ~ –9.90, *p*-value < 0.0001; /ul/: Est ~ –96 ms, *t*-value ~ – 9.95, *p*-value < 0.0001; /aɪr/: –18 ms, *t*-value ~ –3.28, *p*-value < 0 .001; /ir/: –32 ms, *t*-value –4.31, *p*-value <0.0001).

Next, diphthong rimes /aɪl/ and /aɪr/ are significantly longer than all remaining rimes (with averaged rime duration differences ranging from 23 to 56 ms (<0.001 significance levels). For either of the two measures the /aɪl/ and /aɪr/ rimes do not significantly differ from each other: average duration – *t*-value ~ 0.28; *p*-value ~ 0.7; normalized duration – *t*-value ~ 0.35; *p*-value ~ 0.72.

Furthermore, /id/ rimes are longer than /il/ rimes (average duration: ~26 ms longer, *t*-value ~ 3.35, *p*-value ~ 8.7e–04). /il/ and /ir/ rimes do not differ in their contribution to the rime (normalized duration: *p*-value ~ 0.2) or in their averaged duration (*p*-value ~ 0.2).

Finally, for stimuli involving the tense back vowel /u/, rimes with a nasal coda are the longest (e.g., average duration: /un/ is 33 ms longer than /ul/ rimes, *t*-value ~3.49; *p*-value 5.2e–04), and rimes with the stop coda /k/ do not differ from /ul/ rimes (average duration: *p*-value ~ 0.08).

Results show that liquid rimes are not consistently longer than rimes involving stops or nasals. In addition, rhotic rimes are not longer than lateral rimes. None of these duration patterns could predict the syllable counts attributed. Recall that the patterns observed for the SCJ task showed over-one SCJs being attributed exclusively to words with liquid codas. Furthermore, sesquisyllables involving rhotics receive more over-one SCJs than those involving laterals. The two results taken together thus indicate that rime duration alone is not a good predictor of participants' decisions about syllable counts. In the next section we further probe the role of duration by testing the correlation between SCJ responses and rime duration.

3.3. Correlating syllable count judgments and rime duration

Figure 9 illustrates the differences in normalized and averaged rime duration associated to each SCJ answer option as a function of coda_type. A cumulative link mixed model (clmm) with flexible thresholds and logit link function, was run for each measure with SCJ as an ordered response variable, average rime_duration, coda_type, and lexical frequency as fixed factors, and Participant and Repetition as random factors. Random intercepts were included. Likelihood tests (performed on nested models) show there is an effect of rime duration (averaged duration: $\chi^2(1) = 50.105$, $p < 0.0001$; normalized duration: $\chi^2(1) = 10.668$, $p < 0.0001$), coda_type ($\chi^2(1) = 13.605$, $p < 0.0001$) and lexical frequency ($\chi^2(1) = 5.154$, $p < 0.01$).

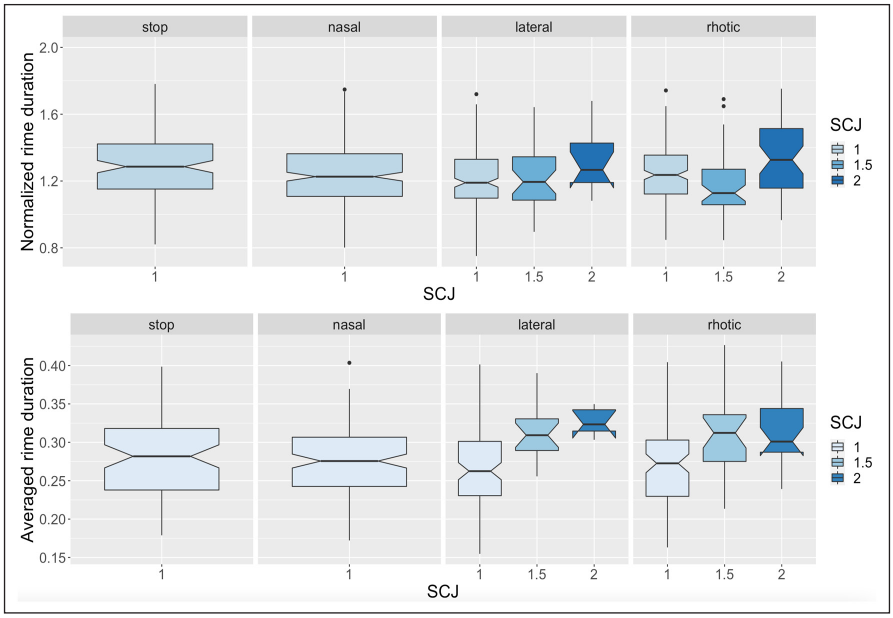

Figure 9: Normalized and averaged rime duration per SCJ answer option (1, 1.5, 2 syllables) and coda type (stop, nasal, lateral, rhotic).

Duration is not a good predictor of SCJs overall, when including the controls. However, within the subset of sesquisyllables, the analysis correlating duration and SCJs shows that averaged rime duration is a good predictor for over-one SCJs. Over-one SCJs are associated with longer rimes: The longer the liquid rime, the higher SCJs are likely to be (Estimate 27.16; z-value $\sim$ 5.75, p-value $<$ 0.0001; normalized duration: 2.01, z-value$\sim$3.315, p-value $<$ 0.0001). Rhotic rimes are more likely to receive over-one SCJs than lateral rimes (averaged duration: Estimate 1.41, z-value $\sim$ 3.75, p-value $<$ 0.0001; normalized duration: 1.05, z-value $\sim$ 4.522, p-value $<$ 0.0001).

In summary, we found that a third of our participants awarded ambiguous, over-one SCJs exclusively to monosyllabic words involving liquid codas, thus confirming the presence of sesquisyllables in American English. Duration was found to be a good predictor for over-one SCJs within sesquisyllables only: Higher SCJs were associated with longer rime durations for liquid rimes, confirming Tilsen & Cohn's (2016) findings. However, results also show that, taking all coda types into account, rime duration alone is not a good predictor. Rimes of similar duration (e.g., /uːk/ in *spook* and /uːl/ in *pool*) do not yield similar SCJs. Furthermore, the longest rime was recorded for a stop coda token (e.g., *page*) which did not receive any over-one SCJs from participants. These results suggest that acoustic rime duration does not play a direct role in speakers' intuitions about syllable counts. Rather, it is one of several interacting factors that together characterize the phonotactic structure of English. The clear delimitation of words with liquid codas from words involving other coda types is consistent with our gestural account hypothesis and justifies pursuing it further.

The next section will describe the production-SCJ experiment for German, designed to further test this hypothesis. Recall that in German, coda laterals do not have sequentially timed gestures. We therefore predict no differences in SCJs between lateral and non-lateral rimes. In other words, a class of sesquisyllables is not expected to emerge in German.

4. Experiment 2: German

Results will be presented following the same structure as those in Experiment 1. For both SCJs and rime duration, a detailed rime identity analysis will be presented followed by correlation results between the two measures.

4.1. Results: Syllable count judgments

Several significant differences can be observed between German and English native speakers' intuitions about syllable counts. In accordance with our predictions, German native speakers did not attribute over-one SCJs predominantly to lateral coda rimes. Over-one SCJs were attributed instead to all types of coda consonants, as well as to open syllables. The absence of an emerging pattern confirms our prediction. Also, contrary to English, two-syllable judgements were attributed only to the unambiguously disyllabic fillers. **Figure 10** shows the counts of attributed SCJ options per coda and nucleus type.

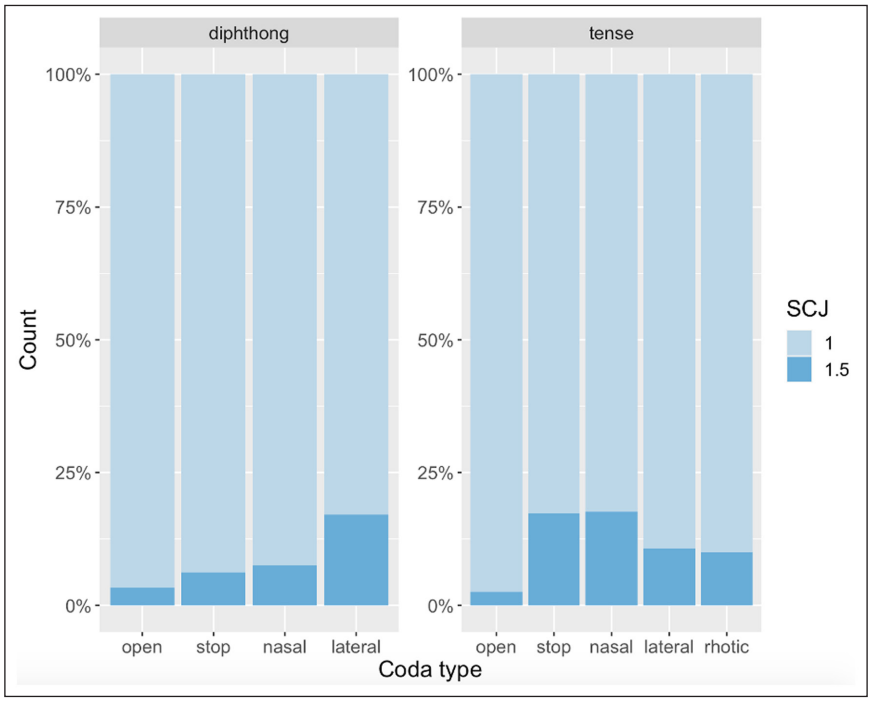

Figure 10: Percentages of attributed SCJs per coda (open, stop, nasal, lateral, rhotic) and nucleus type (diphthong, tense vowel).

When we separate diphthong and tense vowel rimes, we see that diphthong-lateral rimes received the most over-one SCJs, as did tense vowel-stop and tense vowel-nasal rimes. **Figure 11** shows a more detailed count of the SCJs attributed to each rime type.

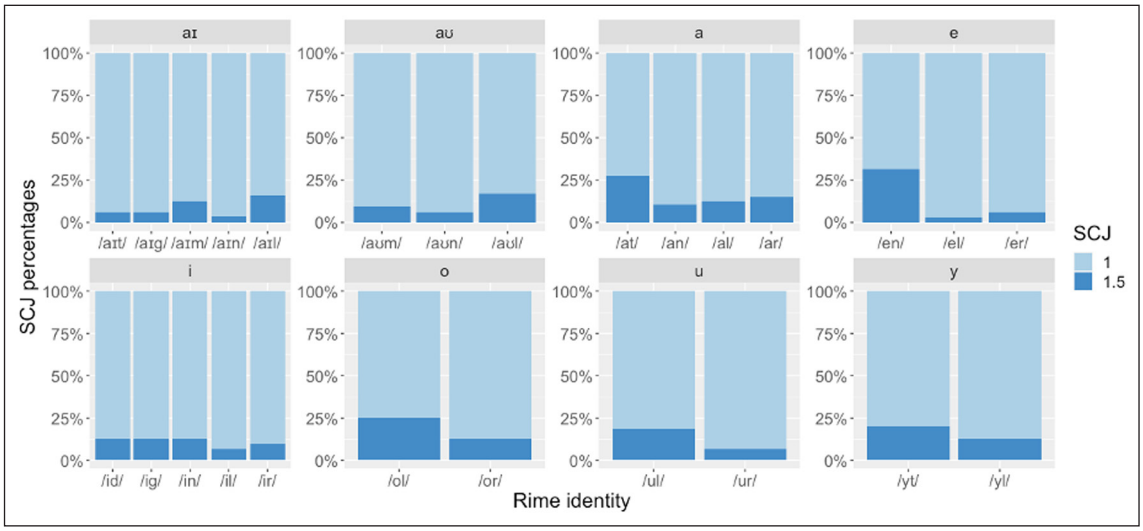

Figure 11: Percentages of attributed SCJs per rime identity and nucleus type. Only CVC tokens included.

Furthermore, all native German participants attributed over-one SCJs, but none are consistent in their choices. **Table 7** shows the percentages of over-one SCJs for target and control tokens as a function of syllable nucleus and rime identity. For identical rime tokens (i.e., Teil, Heil, Seil, feil) most participants gave over-one SCJs randomly to only one of these four /aɪl/ tokens, even though they are all minimal pairs differing in their simple onset consonant.

Nucleus	Rime	token	1.5	2	Total >1
/aː/	/aːn/	Wahn	2	0	2
	/aːl/	Saal	2	0	2
	/aːr/	Haar	3	0	3
	/aːt/	Staat	4	0	3
/eː/	/eːn/	Fehn	6	0	6
	/eːl/	fehl Mehl	0 1	0 0	0 1
	/eːr/	sehr	1	0	1

(Contd.)

Nucleus	Rime	token	1.5	2	Total >1
/iː/	/iːd/	Lied	2	0	2
	/iːg/	stieg	2	0	2
	/iːn/	Wien	2	0	2
	/iːl/	viel	1	0	1
		Stiel	2	0	2
		Kiel	1	0	1
	/iːr/	vier	0	0	0
		Tier	1	0	1
		Stier	2	0	2
		Bier	2	0	2
/oː/	/oːl/	wohl	2	0	2
	/oːr/	Chor	2	0	2
/uː/	/uːl/	Stuhl	4	0	4
	/uːr/	fuhr	1	0	1
/y/	/yːt/	blüht	4	0	4
	/yːl/	kühl	2	0	2
/aɪ/	/aɪt/	seit	1	0	1
	/aɪg/	Teig	1	0	1
	/aɪm/	Heim	2	0	2
	/aɪn/	sein	1	0	1
		mein	1	0	1
	/aɪl/	Teil	4	0	4
		Seil	5	0	5
		Heil	1	0	1
		feil	2	0	2
/aʊ/	/aʊm/	Baum	1	0	1
		Raum	1	0	1
	/aʊn/	Faun	1	0	1
	/aʊl/	faul	3	0	3
		Maul	4	0	4

Table 7: Number of speakers (out of 16) having chosen 1.5 and 2 SCJs per target token, as a function of nucleus type and rime identity.

The recorded response times varied between 0.3 seconds and 7.88 seconds, with an average response time of 1.02 seconds. **Figure 12** shows the raw RT distributions per coda type. **Table 8** shows the mean, median, and standard deviation values per coda type, indicating that, contrary to English, there is no difference based on coda type.

	Open	**stop**	**nasal**	**lateral**	**rhotic**
mean	0.77	0.96	1.09	1.02	1.07
median	0.6	0.76	0.82	0.77	0.83
standard deviation	0.39	0.64	1.10	0.82	0.78

Table 8: Mean, median, and standard deviation for reaction times distributions per coda type.

Table 9 shows the results for the Gamma log link glmm (fixed factors: coda_type and lexical frequency; random factors: Participant with random intercepts). Results show that reaction times involving lateral codas are not significantly different from other coda types (except for no coda tokens). Furthermore, contrary to English, lexical frequency did not play a role for participants' reaction times.

	Measures	**open**	**stop**	**nasal**	**rhotic**
lateral	Estimate	–0.20	–0.09	0.06	0.067
	t-value	–2.58	–0.14	1.01	0.97
	p-value	<0.001	0.88	0.30	0.33

Table 9: Glmm results: significance levels for RT (log values) as a function of coda type.

Contrary to the English data, there are no patterns emerging in the German participants' SCJs. Three possibilities could explain the lack of consistency: (1) speakers attributed over-one SCJs independently of either coda type or rime duration, only because the option was available to them, (2) duration might play a bigger role in German than in English, which we will investigate further, or (3) frequency effects might be at play. The reaction time analysis ruled out an effect of frequency, leaving two options: Either participants attributed SCJs randomly, independently of coda type or rime duration, or rime duration played a more significant role in German than in English. If the latter is true, we would expect diphthong – lateral rimes to be longer compared to other diphthong rimes, and nasal and stop codas to contribute the most to tense vowel rimes.

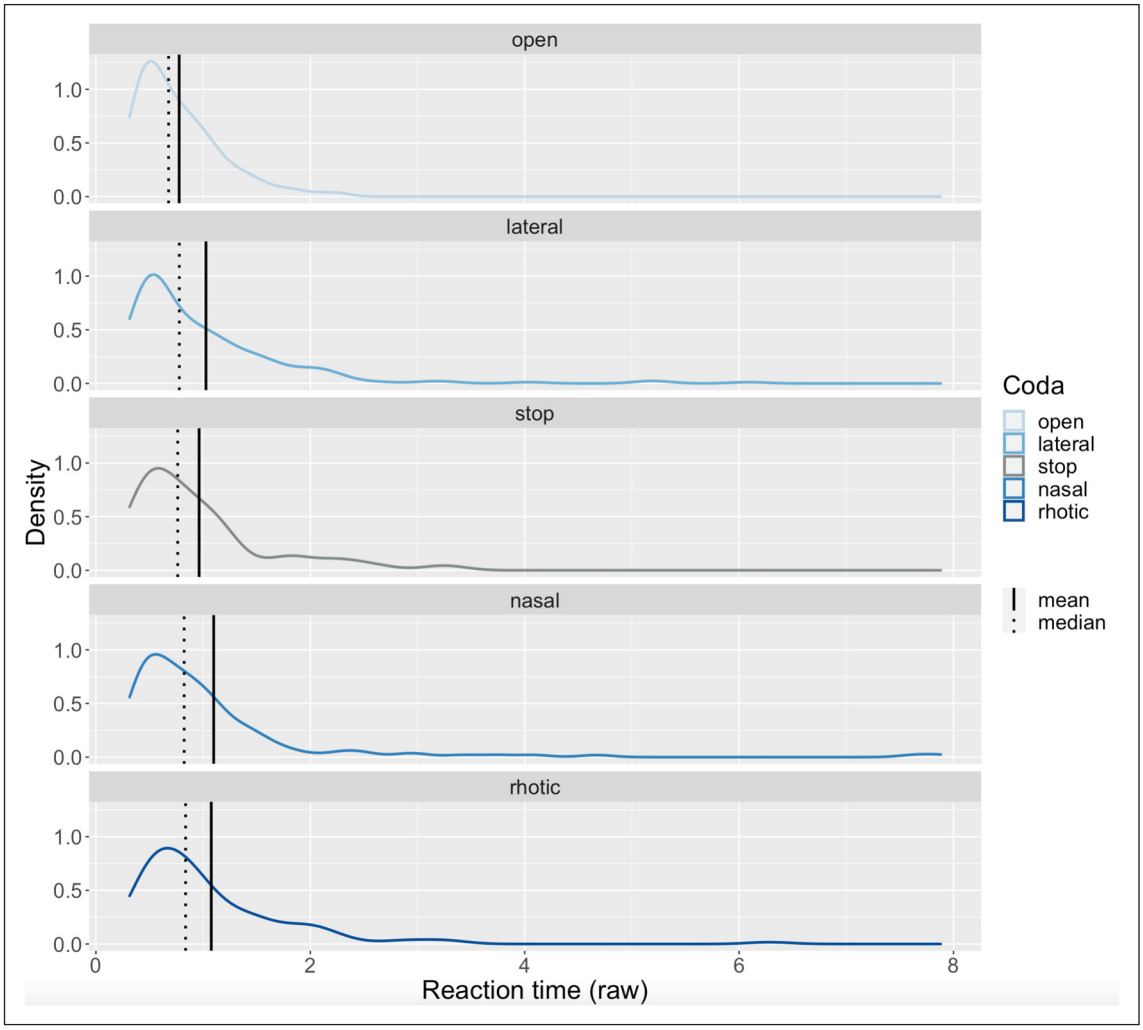

Figure 12: Raw distributions of response times per coda type. Vertical solid lines indicate mean values; vertical dashed lines indicate median values.

4.2. Results: Duration

Duration results will be presented as in Experiment 1. Differences over different classes of rimes in normalized and averaged rime duration, based on rime identity analysis, will be presented. Likelihood ratio tests showed an effect of rime_identity compared to the null model (averaged duration: $\chi^2(22) = 350.42$, $p < 0.0001$; normalized duration: $\chi^2(22) = 350.42$, $p < 0.0001$). As in the case of English, there was no significant effect of lexical frequency on rime duration (averaged duration: $\chi^2(1) = 1.46$; $p \sim 0.23$; normalized duration: $\chi^2(1) = 1.63$; $p \sim 0.31$). The final model included only rime_identity as fixed factor, and Participant and Repetition (only for the normalized duration measure model) as random factors. Random intercepts were included in all models.

Figures 13 and **14** show the normalized and averaged rime duration, respectively, for individual rimes, based on nucleus (facets) and coda type (colors).

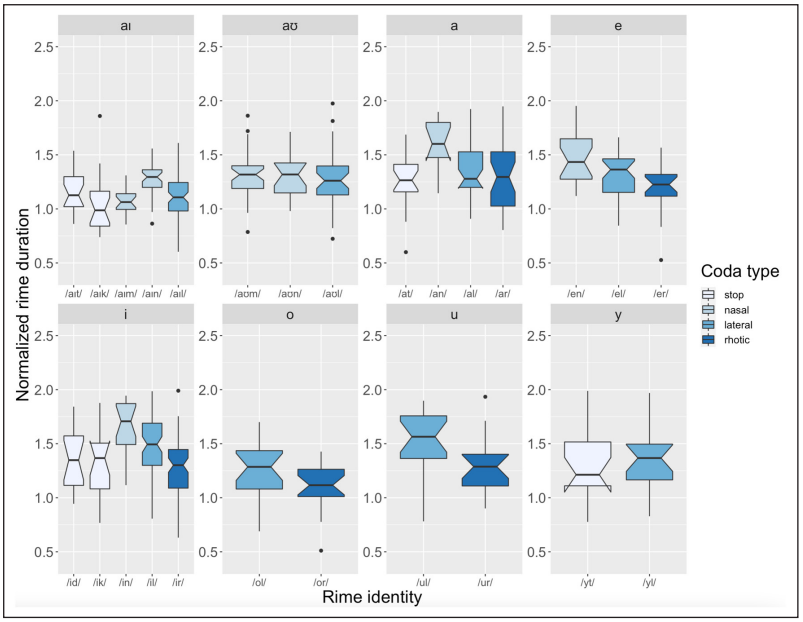

Figure 13: Normalized rime duration per rime identity and nucleus type (aː, aɪ, aʊ, eː, iː, oː, uː) and coda type (open, stop, nasal, lateral, rhotic).

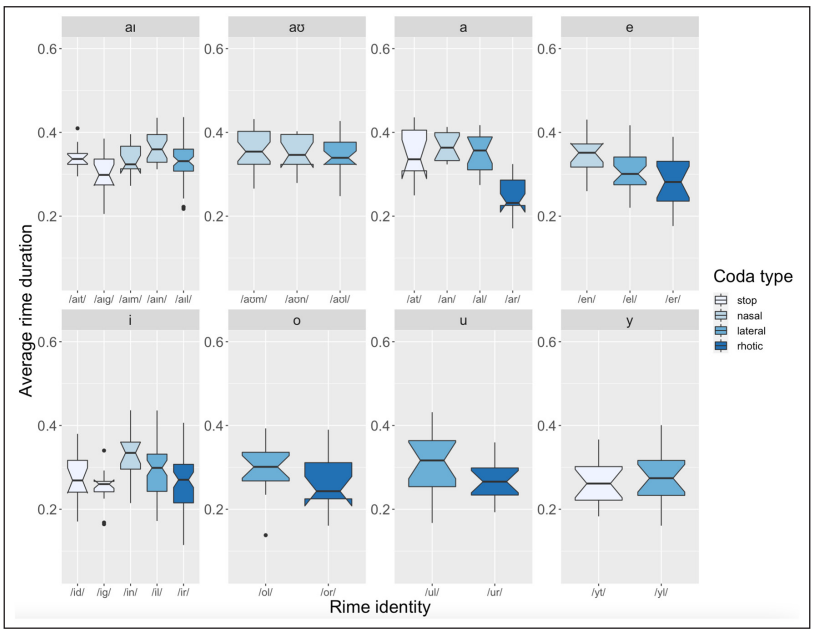

Figure 14: Average rime duration per rime identity and nucleus type (aː, aɪ, aʊ, eː, iː, oː, uː) and coda type (open, stop, nasal, lateral, rhotic).

The results of the full pair-wise comparison (based on *rime_identity*) are given in Appendix C and D. Here we highlight the essential results. For diphthong rimes, there are no significant differences based on coda type for either of the two measurements. Only /aɪn/ rimes are significantly longer (~49 ms, *p*-value ~ 0.02). More variability in duration was found for tense vowel rimes, depending on coda type. For tokens containing the vowel /a:/, /a:l/ rimes are significantly longer than /a:r/ rimes (86 ms, *p*-value ~ 2.47e–07) but shorter than /a:n/ rimes (–49 ms, *p*-value ~ 0.003). No difference between /a:l/ and /a:t/ rimes (*p*-value ~ 0.7) was found. A nasal coda contributes marginally more to rimes involving the tense vowel /e:/ than a lateral coda (*p*-value ~ 0.6), but the average duration of /e:n/ rimes is significantly longer than /e:l/ rimes (46 ms, *p*-value ~ 0.002). The difference in duration between /i:l, /i:n/, and /i:r/ rimes is significant for both duration metrics. No difference in duration for either measure was found for tokens involving the tense vowels /o:/ and /y:/. /u:l/ rimes were however longer than /u:r/ rimes (43 ms, *p*-value ~ 0. 01).

In summary, in German, duration does not play a role in participants' SCJs in the case of diphthong rimes, and only partially correlates with SCJs in the case of tense vowel rimes. While in some cases (e.g., /e:n/), the highest count of over-one SCJs corresponds to the longest rimes, in others it does not. Rimes in /a:r/, for example, receive more over-one SCJ than /a:l/, but rime duration is significantly shorter. There appears to be no straightforward relation between SCJs and rime duration in German, and it will be confirmed by the analysis in the next section.

4.3. Correlating syllable count judgments and duration

Figure 15 shows the difference in rime duration per SCJ option (1 versus 1.5) as a function of coda type. As in Experiment 1, a cumulative link mixed model (clmm) was considered, with logit link function and flexible thresholds, with SCJ as an independent variable, rime_duration, coda_type, and lexical_frequency as fixed factors, and Participant as a random factor. Open syllables have been excluded because there were too few observations of 1.5 SCJs, which resulted in an ill-fitted model (high condition number of the Hessian). Likelihood ratio tests on nested models show no effect of rime_duration ($\chi^2(1) = 0.007$, $p \sim 0.9$), or coda_type ($\chi^2(3) = 1.74$, $p \sim 0.6$), or lexical_frequency ($\chi^2(3) = 0.74$, $p \sim 0.4$).

In summary, we found that there is no straightforward relation between SCJs and rime duration in German. Contrary to English, German native speakers attributed the 1.5 syllable count apparently randomly, independently of coda type, duration, or lexical frequency, to words involving all coda types, and they never chose the 2-syllable option. Duration was found to be only partially correlated with SCJs. No duration difference was found between tokens that received 1 versus 1.5 syllable counts. The lack of correlation between SCJs and duration, as well as the absence of a specific pattern, confirm the absence of sesquisyllables in German, supporting our hypothesis, and justifying the search for an alternative, gestural account of sesquisyllables.

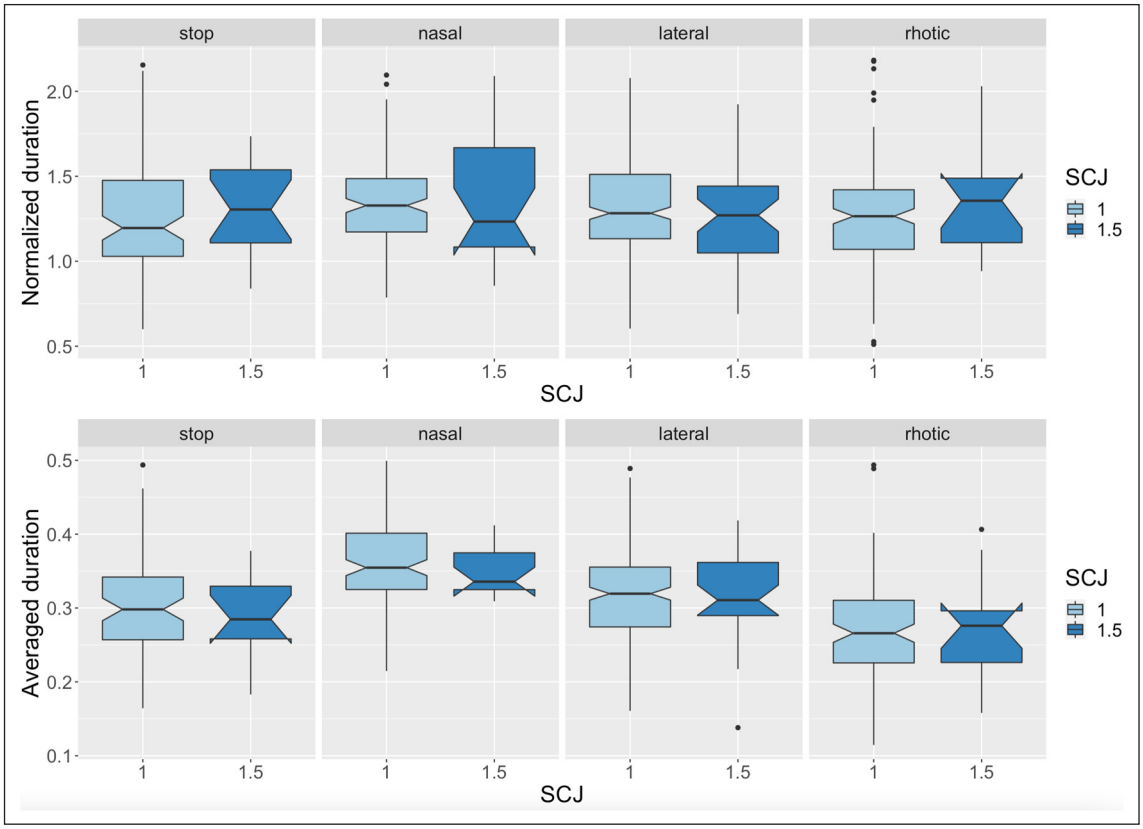

Figure 15: Average rime duration per SCJ answer option (1 and 1.5 syllables) and coda type (stop, nasal, lateral, rhotic).

5. Proposed analysis

Before discussing the implications of the results for phonological representation, we will summarize the main findings of our experiments. Results from experiments 1 and 2 are threefold, and support our hypothesis and predictions presented in Section 1.2.

5.1. Main findings

First, sesquisyllables are confirmed as a special class of words in English, but not in German. In English, over-one SCJs are attributed exclusively to rimes involving diphthongs or tense vowels followed by a liquid consonant, thus differentiating liquid from non-liquid codas. In German, participants attributed over-one SCJs randomly across both targets and controls, independently of coda type. English and German native speakers also differ in the consistency of their attribution of ambiguous over-one SCJs. In English, three participants attributed over-one SCJs consistently to targets involving most liquid coda tokens. Three more participants extended over-one SCJs to diphthongs followed by a liquid coda, and the remaining 12 participants all attributed over-

one SCJs at least once, always to words containing a diphthong plus a liquid coda. In German, participants were inconsistent—all 16 participants attributed over-one SCJs at one point during the task, but to a random variety of rimes.

Response times also seem to support a differentiation between liquid and non-liquid codas in English, but not in German. Native English speakers took a significantly longer time to attribute SCJs to stimuli involving liquid than non-liquid codas. Therefore, even when participants ended up attributing just one SCJ to stimuli with liquid codas, they took a longer time to choose their answer, and this could be interpreted as an indication of sesquisyllables. In German, no difference based on coda type was found for closed syllable stimuli, indicating that monosyllabic words with liquid codas are not treated differently from their non-liquid counterparts.

Second, acoustic rime duration across all rime types does not correlate with participants' intuitions about syllable counts, in either English or German. In English, rimes with liquid codas are not longer than those with non-liquid codas, but they are the only ones that receive over-one SCJs. In German, duration only partially accounts for the attributed SCJs – the longest rimes do not always receive the highest syllable counts. This result is also confirmed by the model correlating SCJs and duration.

Third, in English only, rime duration within the class of sesquisyllables does correlate with participants' SCJs. For words with liquid codas only, non-monosyllabic judgments correspond to longer rime durations. Furthermore, within the subset of sesquisyllables, over-one syllable counts were more likely to be attributed to tokens involving rhotics and to words that were less frequent. In German, where no effect of rime duration, coda type, or lexical frequency was found for either duration metric, both monosyllabic and over-one syllable counts were attributed to rimes of similar duration, independent of coda type.

In summary, results are consistent with the presence of sesquisyllables in English, and with their absence in German. Acoustic duration being a by-product of articulatory synergies within the rime, in the next section we discuss articulatory accounts of sesquisyllables, and we develop our proposal.

5.2. Gestural representation of sesquisyllables

The main hypothesis of the present study is that syllable count judgments are linked to the phonological representation of articulatory timing patterns. We focus on the representation of gestures as abstract subphonemic units, and on their temporal and spatial coordination, following Articulatory Phonology (Browman & Goldstein, 1986; 1990; Goldstein & Fowler, 2003; Saltzman & Munhall, 1989). In Articulatory Phonology, the cognitive and the physical dimensions of speech are integrated into one complex self-organizing system. Gestures are discrete actions of the speech articulators. They form and release phonologically relevant constrictions, and are

defined by inherent spatial properties (constriction location and constriction degree), as well as by temporal properties. Gestures are active for a given interval, and are timed with respect to each other. For the representation of liquids in sesquisyllables, we focus specifically on the timing of the gestures within a segment and within a syllable.

We hypothesized that the presence of a special category of sesquisyllabic words in English is related to the inherent gestural complexity of liquid coda consonants in this language. More specifically, we argue that the temporal coordination and the quality of the dual gestures involved in the production of laterals and rhotics in English, give rise to ambiguous syllable counts among native speakers. By comparing English to German, a language with similar word structure and vowel length contrasts, but different articulation for the lateral coda, we isolated the gestural parameters that allow us to test our hypothesis.

The temporal coordination of coda liquids in English has been amply studied (Sproat & Fujimura, 1993; Giles & Moll, 1975; Alwan et al., 1997; Narayanan et al., 1997; Browman & Goldstein, 1995; Gick, 1999; Marin & Pouplier, 2010; 2014; Proctor et al., 2019). The specificity of English coda liquids as opposed to the onset concerns the sequential timing of two gestures: The tongue dorsum (TD) gesture precedes the tongue tip gesture, for both lateral and rhotic codas. Furthermore, the rhotic and lateral TD gestures are considered to be of vocalic quality, because of their slower speed of constriction formation and release, more similar to that of vowel gestures.

Hence our proposed model entails that in sesquisyllables, the vocalic TD gesture of the liquid coda follows two vocalic gestures, which are those of a diphthong or of a tense vowel nucleus. The resulting structure thus contains three sequentially coordinated vocalic gestures. We argue that this particular structure (i.e., three sequentially coordinated vocalic gestures) receives ambiguous over-one SCJs. Beyond this particular gestural configuration, in words involving (1) lax vowels followed by any coda consonant, or (2) tense vowels/diphthongs followed by a non-liquid consonant, the third sequential vocalic gesture does not exist, and only classic monosyllabic judgments are obtained.

Recall that in German, the lateral coda consonant does not involve a TD retraction preceding the TT gesture. While conclusive articulatory evidence is still missing on whether the German lateral has a target specification for its TD component or not, we assume for now that the German /l/ does not have a TD retraction preceding the TT gesture (Recasens et al., 1998). In German then, tense vowels and diphthongs are followed by a liquid consonant that does not exhibit a sequential relation between the two gestures. A third vocalic gesture in the sequence is never present, and no sesquisyllables exist.

In the rest of this section, we first introduce existing models of liquid consonants in American English rimes, closely related to the model we propose. We then develop our model within the Articulatory Phonology framework. In Section 6 we discuss the implications of our model and we suggest future research avenues that can further test it.

5.3. Articulatory account of American English liquid rimes

The articulatory specificities of American English coda liquids have been addressed within the framework of Articulatory Phonology (Browman & Goldstein, 1986; 1990; Saltzman & Munhall, 1989), and the coupled oscillator model (Nam et al., 2009; Goldstein, Nam, Saltzman, & Chitoran, 2009), which defines syllables in terms of gestural coordination patterns. Each gesture within the syllable is associated with a nonlinear coupled oscillator, and its activation is triggered at a particular phase of its oscillation. Gestures in the onset are hypothesized to be coupled in-phase (synchronously) to the vowel gesture of the nucleus, while in the rime the gestures are coupled in an out-of-phase coordination (Turvey, 1990), resulting in a sequential activation of the gestures in the vocalic nucleus and the coda consonant. Derivations of this framework have been used to explain various phonological phenomena, two of which are of interest here: (1) Tilsen and Cohn's (2016) model accounting for duration patterns observed for sesquisyllables and (2) Walker and Proctor's (2019) model accounting for tense/lax vowel neutralization before certain types of liquid rimes in American English. Both models rely on the coordination patterns of subphonemic units in the rime, albeit differently. The common element these models share is the overlap of the gestures involved in the production of the vowel and the coda liquid. More specifically, for Tilsen and Cohn (2016), this concerns the articulatory regimes involved in the overlap of these gestures, while for Walker and Proctor (2019), it concerns the degree of gestural overlap – coarticulation degree. Both models relate gestural overlap to syllabic weight (moraic representation), which in turn provides a basis for the representation of sesquisyllables in American English.

Tilsen and Cohn (2016) rely on Selection-Coordination theory (Tilsen, 2014; 2016), derived from Articulatory Phonology, to provide a theoretical framework for explaining sesquisyllables. The authors argue that Selection-Coordination theory best explains the categorical patterns involved in syllable count judgments by allowing for more gradient variation in gestural overlap. Selection Coordination theory provides two control options for the timing of coda gestures: competitive control and coordinative control. In the competitive control regime, gestures are selected with mutual exclusion, resulting in reduced overlap between them; in the coordinative control regime, gestures can be co-selected, allowing for more overlap between them. Within this framework, variable syllable counts are the result of the speaker's choice of articulatory control regime. The choice of competitive control limits gestural overlap, resulting in more over-one SCJs. Tilsen and Cohn (2016) further relate competitive control to trimoraic structure. Coordinative control, however, allows for gradients of overlap between gestures in the rime, resulting in monosyllabic syllable counts. Coordinative control is thus associated with bimoraic structure.

Walker and Proctor (2019) define gestural overlap in terms of blending strength of subsegments, corresponding to the degree of coarticulatory dominance (Saltzman & Munhall, 1989; Fowler & Saltzman, 1993). Greater coarticulatory dominance corresponds to a stronger specified blending strength, i.e., less overlap. In this model the blending strength is segment-specific, based on evidence from articulatory studies of American English liquids (Proctor et al.,

2019), which show that laterals have lower coarticulatory dominance than rhotics, resulting in more overlap between the gestures of the lateral and the nuclear vowel. Within this account, the sequencing and overlap of the gestures are directly related to the moraic count of the rimes.

The purpose of the present study was to test sesquisyllabicity cross-linguistically and propose a model within the Articulatory Phonology framework. The model we present retains from the previous two models the relevance of gestural overlap, and considers an additional factor—the vocalic quality of the gestures involved. We propose a purely gestural model, based on the degree of overlap of the vocalic subphonemic units present in the rime, allowing for gradient representations that can account for speakers' gradient intuitions.

5.4. Proposed model

The model we are proposing introduces a novel factor—the quality of the gestures involved (vocalic versus consonantal). This additional factor allows us to account for the cross-linguistic distribution of sesquisyllables, as well as for the generalizations captured by moraic representations.

Our analysis assumes an interpretation of gestural quality and overlap as they are defined within the Task Dynamics Model of Saltzman and Munhall (1989). Both dimensions are adjustable by changing the model parameters: The stiffness parameter k can be adjusted to fit vocalic or consonantal quality; bonding strength, relative phasing between pairs of gestures, and/or activation windows of gestures can be adjusted to vary gestural overlap.

The count of vocalic gestures in the rime is directly linked to the gestural composition of the nucleus and the coda consonant, to the degree of overlap (controlled by bonding strength, activation windows of gestures, and/or phasing relations) between the vocalic gestures in the rime, and to timing relations between the gestural units of the coda consonant. As described in Section 1.1, coda liquids involve two gestures, one of which is vocalic: a tongue dorsum (TD) retraction gesture in the uvular region for the coda lateral, and a pharyngeal constriction of the TD for the coda rhotic. The second, consonantal gesture, is an alveolar constriction for the lateral, and a palatal constriction for the rhotic.

Diphthong nuclei are composed of two distinct vocalic gestures, with different constriction locations (Kent & Moll, 1972). The gestural composition of tense vowels has been much less studied. Acoustically, tense vowels are defined by multiple parameters (Stevens, 1998), leading to different proposals in terms of featural representation. There is, nevertheless, general agreement that the tense/lax contrast involves a length distinction (Jakobson, Fant, & Halle, 1952, p. 36–39; Stevens, 1998, p. 294–299; Hammond, 1999). Tense vowels also involve movement of the tongue body during the course of the vowel, which often tends toward diphthongization (Stevens, 1998, p. 296–298). We adopt here the only gestural representation of English tense vowels currently available, namely the one proposed in the task dynamic synthesizer TADA (Nam et al., 2004),

and which consists of two vocalic gestures of similar constriction location. The English high tense vowels are defined in TADA as consisting of a nucleus /i/ or /u/, followed by a coda glide /j/, /w/ with similar constriction location. This dual gesture representation of tense vowels is very similar to that of diphthongs, except that the constriction locations of the two vowel gestures are different for diphthongs and similar for tense vowels. While the gestural organization of vowels remains to be further studied, we make, for now, a crucial assumption. We assume that gestures with similar constriction locations, such as those of tense vowels, overlap more than gestures with different constriction locations, as those of diphthongs.

To summarize, our model makes the following assumptions:

1. Diphthongs are composed of two vocalic gestures of different constriction location.
2. Diphthongized tense vowels are composed of two vocalic gestures of similar constriction location.
3. Gestures with similar constriction locations have a higher degree of overlap.
4. TD gestures (the gestures of diphthongs and tense vowels, and the TD gestures of the liquids) are vocalic in nature.
5. The TD gesture of coda liquids (rhotics *and* laterals) are sequentially coordinated to the gesture of the nuclear vowel, either out-of-phase (180°) or at an eccentric phase.

Given these assumptions we can formulate the premise of the model:

> Model premise: SCJs of more than one syllable are attributed to any rime which has more than two sequentially coordinated vocalic gestures.

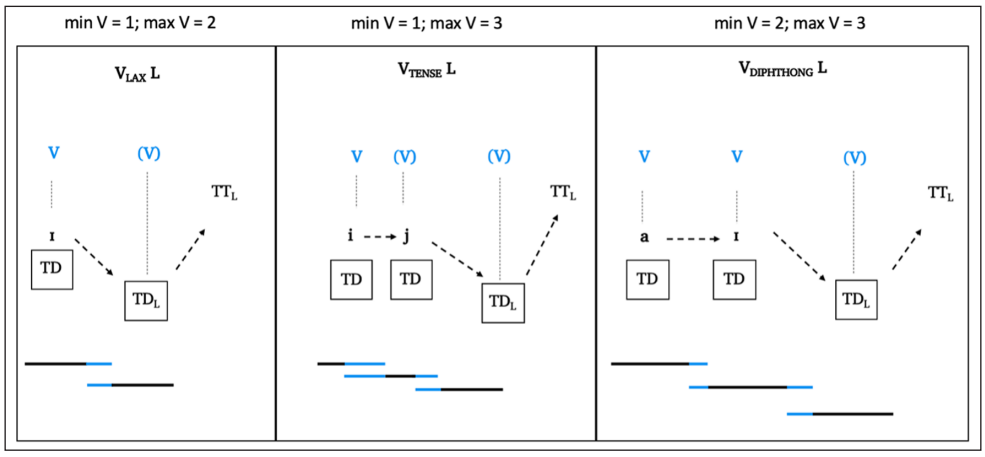

Figure 16: English. Gestural representation and vocalic gesture structure of English rimes involving liquids with sequentially timed vocalic gestures. V in parenthesis denotes optional counts of vocalic gestures in native speakers' SCJs.

Figure 16 illustrates the possible counts of vocalic gestures in rimes with independently timed vocalic TD gestures (in square frames). Underneath each gestural representation are schematic representations of gestural overlap of the vocalic gestures: The blue lines indicate a variable degree of overlap. Depending on the degree of overlap, the vocalic gesture is then counted or not. Reduced overlap results in over-one SCJs.

The leftmost panel illustrates the case of a lax vowel rime (e.g. *fill, pull*). The lax vowel accounts for one vocalic gesture. The vocalic TD gesture of the coda liquid is activated after the nuclear vowel gesture and before the TT_L gesture of the liquid. The TD_L gesture of the liquid could potentially be counted as a second vocalic gesture in the rime, depending on the degree of overlap between the vowel TD and the TD_L gestures. Based on assumption (3) we would expect more overlap between the vowel and TD_L in the case of *pull* than in the case of *fill*. In either case, however, the minimum vocalic gesture count is 1, and the maximum vocalic gesture count for a lax vowel – liquid rime is 2, which would predict only monosyllabic judgments.

The middle panel illustrates the case of tense vowel – liquid rime (e.g., *feel, pool, beer*). Given assumption (2), the tense vowel nucleus can count for one or two vowel gestures in SCJs, depending on the degree of overlap of the two vocalic gestures. The TD_L gesture of the liquid would then either count as a second vocalic gesture or as a third one, depending on the degree of overlap with the preceding vowel. The minimum vocalic gesture sequence is thus 1 and the maximum vocalic gesture count for tense vowel – liquid rimes would be 3. The former predicts monosyllabic syllable counts, the latter over-one syllable counts. A sequence of 2 vocalic gestures would result in a monosyllabic judgment.

Finally, for the case of diphthong – liquid rimes (e.g., *file, pile, fire, pyre*), the diphthong corresponds to two distinct vocalic gestures and the TD_L gesture could count for a third vocalic gesture, depending on its degree of overlap with the preceding one. In diphthong – liquid rimes, therefore, the minimum vocalic gesture sequence is 2, and the maximum count is of 3 vocalic gestures, predicting monosyllabic and over-one SCJs, respectively.

So far, we considered only the case of coda liquids for which the vocalic TD gesture occurs before the consonantal tongue tip gesture. In German, the lateral coda differs in the timing and the constriction location of the vocalic gesture: The TD gesture, if at all present, is lowered, not retracted, and is synchronous to the tongue tip gesture. Because of the synchronicity of the two gestures (or the lack of TD_L gesture), we assume that the TD_L gesture is never counted as a vocalic gesture in SCJs, and a tri-vocalic sequence cannot occur. This is illustrated in **Figure 17**. The vocalic gesture count in clear lateral rimes in German is thus maximally 2, corresponding to monosyllabic judgments across the board.

Given the hypothesized gestural configuration of German clear /l/ and our model premise—that over-one SCJs are attributed to any rime which has more than two sequentially coordinated vocalic gestures—sesquisyllables are not predicted to occur in German.

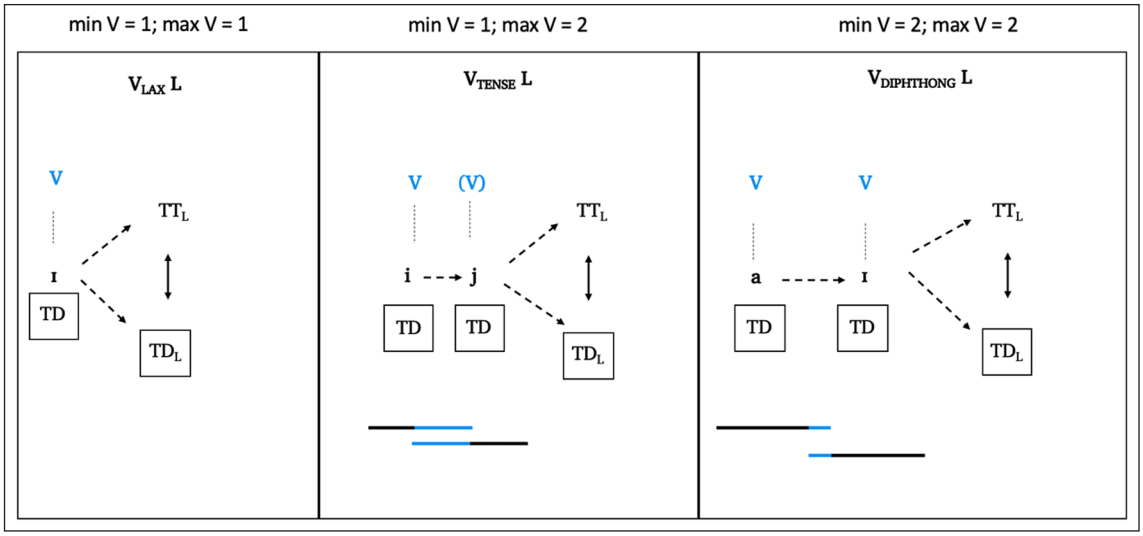

Figure 17: German. Gestural representation and vocalic gesture structure for ɪl (left), iːl (center), and aɪl (right) rimes.

The model, as described above, can predict the presence of sesquisyllables in English and their absence in German. The next section discusses the implications of the model for the observed SCJ patterns studied here, and for other languages, as well as the limitations of our present study. We will end with a general conclusion.

6. Discussion

6.1. Implications

The proposed model accounts for the presence of sesquisyllables in English and their absence in German by relying on the quality (vocalic versus consonantal), the timing, and the degree of gestural overlap of subphonemic units composing the syllable rime. The interaction between the degree of overlap and the quality of the gestures involved accounts for speaker-specific and word-specific variation.

Two main patterns were observed for participants' SCJs in English: (1) sesquisyllables containing diphthongs received more over-one SCJs than those containing tense vowels, and (2) sesquisyllables with rhotic codas received more over-one SCJs than those with lateral codas. The first pattern is accounted for by the assumption that similar sequential vowel gestures, as in tense vowels, have a higher degree of overlap. While diphthongs are composed of two different vowel gestures, exhibiting less overlap, tense vowels involve two similar vocalic gestures which can vary in overlap. They can have more overlap than diphthongs, or as little overlap as the different gestures of diphthongs. The model predicts that words with more overlapped vowel gestures are unlikely to receive over-one SCJs. Tense vowels with two less overlapped vocalic gestures

are more likely to receive over-one SCJs, provided that the degree of overlap between the two gestures is sufficiently reduced, similar to the one in diphthongs.

The second SCJ pattern observed, opposing rhotic to lateral codas in sesquisyllables can also be accounted for by different degrees of overlap, more specifically by the difference in patterns of coarticulation between laterals and rhotics. In an extensive real-time MRI study, Proctor et al. (2019) found that rhotics are more resistant to coarticulation and more coarticulatory aggressive than laterals, especially in coda position. Both the coarticulatory resistance and aggressiveness depend on the articulators involved and the manner of articulation. The articulator—the tongue dorsum—is the same for both liquids (Proctor et al., 2019; Alwan et al., l997; Narayanan et al., 1997; Espy-Wilson et al., 2000) but in the case of rhotics their vocalic gesture is less prone to overlap with the preceding vowel. In our model, this would result in higher counts of vocalic gestures for rimes with a rhotic coda, yielding more over-one SCJs for coda rhotics than for coda laterals.

Interestingly, the same patterns were confirmed for speakers of non-rhotic British English (Popescu, 2019). In these dialects words that are spelled with final 'r' have either tense vowels or schwa in the rime (Heffner, 1950). This result additionally supports our hypothesis that sesquisyllables are linked to the count of vocalic gestures in the rime—three sequentially coordinated vocalic gestures yield over-one SCJs—and that the consonantal gesture plays only a marginal role in this respect. In certain English dialects (e.g., Southern British English), laterals are also subject to loss of consonantal tongue tip gestures—coda vocalized /l/ exhibits a reduced or altogether absent coronal constriction gesture (Strycharczuk, Derrik & Shaw, 2020). The case of /l/-vocalization would, within our model, still predict sesquisyllables involving lateral codas.

In the next section we discuss the limitations of our study and of our proposed model, which will lead to a more general discussion about phonological, and in particular, gestural representation.

6.2. Limitations and future research avenues

We acknowledge several limitations to our study and model. A first methodological limitation concerns the reliability of syllable count judgements, and their relationship with phonological representation. It is known that factors such as language, spelling, lexical frequency, or the choice of syllabification task can all contribute to speakers' metalinguistic judgments (Derwing & Eddington, 2014; Côté & Kharlamov, 2011). One cannot definitively say how SCJs are attributed or whether they relate to any phonetic factors at all, even less so if the task is counter-intuitive (i.e., matching non-canonical SCJs to canonically monosyllabic words). However, there is an important mitigating factor: The same SCJ task was presented to both English and German native participants. The emergence of a clear pattern in English and of a seemingly random one in

German is evident, and indicates that, in this specific context, SCJs can still be informative about the hypothesized link between metalinguistic knowledge and articulatory representation.

A second limitation of our study is that it is solely based on acoustic data, while the predictions are rooted in articulatory specifications. To fully test the proposed model, we need to correlate English and German native speakers' actual articulation of coda liquids to their corresponding SCJs. Articulatory data will allow more detailed preliminary observations. As far as we know, the timing of the TD and TT gestures of the German coda lateral has not yet been the main focus of an articulatory investigation. Therefore, an articulatory study looking specifically at the intra-segmental timing and inter-speaker variation of this timing is needed to make better predictions. Furthermore, research on the gestural composition of vowels is scarce. In making assumptions for our model we rely exclusively on the representation of American-English tense vowels proposed in TaDA. But without empirical verification, these representations are very limiting.

Kinematic data can also shed light on our finding in Experiment 1, that even though duration does not correlate with SCJ patterns across all rime types, within the subset of sesquisyllables, over-one SCJs are attributed to longer rimes involving liquids. Yuan and Lieberman (2009) found a positive correlation between /l/ duration and /l/ darkness for syllable final /l/ preceding a boundary. Darker /l/s are characterized by more retraction of the tongue dorsum. Hence a more retracted slower vocalic gesture resulting in longer /l/ duration could explain the correlation between SCJ and rime duration we found in Experiment 1. Variations in acoustic duration, however, could also be the result of intergestural phasing (Browman & Goldstein, 1990), which results in different degrees of overlap between the vocalic gestures. In order to fully understand the link between gestural composition of the rime and SCJs, measures of relative phasing and coupling strength need to be correlated with SCJs in addition to the stiffness parameter.

A theoretical limitation of our model stems from its restriction to sesquisyllables. Ideally, the model should account for other phonotactic patterns. For example, in its current state, it cannot predict the illegality of monosyllabic words with lax vowels and no coda (i.e., */pɪ/, */tʊ/), or of words with tense vowel/diphthong followed by a coda cluster consisting of a liquid and a non-coronal consonant (*/miːłk/, */faɪrg/). Moraic theory handles this problem by imposing bounds on lower and upper moraic counts for words, but requires adjustments pertaining to mora attribution to coda consonants like coerced weight or the extra-syllabicity of coronal consonants. These questions need to be further explored, as they address the very important question of the nature of phonological representations, particularly with respect to the role of the mora.

In the same context, as a further challenge, it is worth taking the theoretical question and the empirical investigation back to the Southeast Asian languages for which the term 'sesquisyllable' was originally proposed (Matisoff, 1973; 1989), to indicate words of one syllable 'and a half' (for a recent review, see Brunelle et al., 2020). Although the specific word shapes concerned are quite different from the words in our study, the relevance of gestural representations can benefit from

being confronted with such data. Butler (2014) is a notable study in this direction, proposing a gestural account of sesquisyllables in three SE Asian languages.

Beyond the empirical observation of the syllable count, a very important issue is that of the variability of judgments. In light of our hypothesis, explaining inter-speaker variability in syllable count judgments by variability in representation is the biggest challenge. This is where the English sesquisyllables crucially differ from the Southeast Asian sesquisyllabic word shapes. The latter are 'canonical' (Brunelle et al., 2020, p. 346), which presumably means that all native speakers share the same unambiguous phonological representations. But the English case, especially in its comparison with German, highlights the issue of variability and prompts the question of typological predictions. What predictions does our model make about other languages?

According to our model, sesquisyllables in English are a special structure that emerged from the combination of specific types of vocalic nuclei and specific types of dual-gesture codas. Both conditions are necessary for the sequential coordination that we propose to affect syllable count judgments. Do other languages meet these conditions? Such languages would have a vowel length contrast or a tense/lax contrast like English, and a sequential coordination of gestures in coda liquids. A possible language to consider is Dutch. Dutch dialects that are particularly interesting are those that have a vowel length contrast and dark /l/ in coda position, but do not have a syllable position allophony for laterals. If the lateral is dark in both positions, should we expect to find variable syllable count judgments and sesquisyllabic responses among native speakers? Not necessarily. We believe that in addition to the sequential coordination of more than two vocalic gestures, the asymmetry of positional allophony may be necessary to destabilize categorical judgments. If the coupled oscillator hypothesis is correct, namely that what defines the syllable is sequential out-of-phase coordination on the right, coexisting with in-phase coordination on the left, then syllable count judgments may shift if the asymmetry is enhanced. Thus, a clear lateral in the onset and a dark one in the coda, accumulating sequentially coordinated vocalic gestures on the right, exaggerates the asymmetry, and may prompt a shift in SCJs. But a dark /l/ in both onset and coda positions is a balanced structure, which would presumably be harder to destabilize enough to lead to a shift in intuitions. Testing for variable sesquisyllabic judgments in relevant dialects of Dutch would thus be an informative test for our gestural account, regardless of whether their presence can be confirmed or not.

7. Conclusion

The goal of our study was to propose a gestural account of sesquisyllables within the Articulatory Phonology framework and test it in a cross-language comparison. Correlations between acoustic rime duration and syllable count judgments in English and German, two languages with vowel length contrast but different gestural specifications for coda liquids, confirmed that sesquisyllables exist in English but not in German. We hypothesized that this discrepancy is primarily related to

the presence of an earlier occurring vocalic (TD) gesture inherent to English coda liquids, and its synchronicity in German. Results of the acoustic analysis of rime duration and its correlation with syllable count judgments in the two languages support our hypothesis. Acoustic rime duration was found to partially correlate with syllable count judgments: Higher syllable count judgments were attributed to longer rimes only within the subclass of sesquisyllables, suggesting duration as a result of intra- and inter-segmental coordination patterns is correlated with syllable count judgments. A model accounting for the observed syllable count judgments, based on gestural representations, was proposed. While articulatory data is needed to reliably test the proposed model, studies matching acoustic data with SCJs are crucial in providing precise hypotheses and predictions. Articulatory data, analyzing the exact relationship between gestural parameters, intergestural timing, and SCJs will be a step forward to finding a unifying account for different phonological phenomena.

Additional files

The additional files for this article can be found as follows:

- **Appendix A.** Filler items for Experiment 1 and 2. DOI: https://doi.org/10.16995/labphon.7681.s1
- **Appendix B.** Results of the linear mixed model – Experiment 1 (English). DOI: https://doi.org/10.16995/labphon.7681.s2
- **Appendix C.** Results of the linear mixed model – averaged rime duration (Experiment 2 – German). DOI: https://doi.org/10.16995/labphon.7681.s3
- **Appendix D.** Results of the linear mixed model – normalized rime duration (Experiment 2 – German). DOI: https://doi.org/10.16995/labphon.7681.s4

Acknowledgements

Research supported by the DAAD (DAAD P.R.I.M.E.), the Labex EFL (ANR-10-LABX-0083-LabEx EFL), the Région Île-de-France, and the Bureau des Relations Internationales, Université Paris Diderot. We thank Adamantios Gafos, Louis Goldstein, Doris Mücke, Sam Tilsen, Abby Cohn, and audiences at the University of Potsdam, the Institute of Phonetics, LMU, Munich, and the Chicago Linguistics Society (CLS 54) for insightful comments and feedback. We also thank the anonymous reviewers, whose comments have greatly improved our paper, the student participants, and the Linguistics Departments at the University of Potsdam and University of Chicago, particularly Alan Yu, for generously extending the use of their recording facilities.

Competing interests

The authors have no competing interests to declare.

References

Alwan, A., Narayanan, S., & Haker, K. (1997). Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part II. The rhotics. *The Journal of the Acoustical Society of America*, *101*, 1078–1089. DOI: https://doi.org/10.1121/1.417972

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. DOI: https://doi.org/10.18637/jss.v067.i01

Boersma, P., & Weenink, D. (2019). Praat: Doing phonetics by computer [Computer program]. Version 6.1.53, retrieved 8 September 2021 from http://www.praat.org/

Browman, C., & Goldstein, L. (1986). Towards an articulatory phonology. *Phonology Yearbook*, *3*, 219–225. DOI: https://doi.org/10.1017/S0952675700000658

Browman, C., & Goldstein, L. (1988). Some Notes on Syllable Structure in Articulatory Phonology. *Phonetica*, *45*, 140–155. DOI: https://doi.org/10.1159/000261823

Browman, C. P., & Goldstein, L. (1990). Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics*, *18*, 299–320. DOI: https://doi.org/10.1016/S0095-4470(19)30376-6

Browman, C. P., & Goldstein, L. (1995). Gestural syllable position effects in American English. In F. Bell-Berti & L. J. Raphael (Eds.), *Producing Speech: Contemporary Issues*. For Katherine Safford Harris. AIP Press: Woodbury, NY.

Brunelle, M., Kirby, J., Michaud, A., & Watkins, J. (2020). Prosodic systems: Mainland Southeast Asia. In C. Gussenhoven and A. Chen (Eds.), *Oxford Handbook of Language Prosody*. (pp. 344–354). Oxford University Press. DOI: https://doi.org/10.1093/oxfordhb/9780198832232.013.22

Burt, L., Holm, A., & Dodd, B. (1999). Phonological awareness skills of 4-year-old British children: An assessment and developmental data. *International Journal of Language & Communication Disorders*, *34*(3), 311–335, DOI: https://doi.org/10.1080/136828299247432

Butler, B. A. (2014). Deconstructing the Southeast Asian Sesquisyllable: A Gestural Account. Doctoral dissertation, Cornell University.

Byrd, D. (1995). C-Centers Revisited. *Phonetica*, *52*, 285–306. DOI: https://doi.org/10.1159/000262183

Campbell, F., Gick, B., Wilson, I., & Batikiotis-Bateson, E. (2010). Spatial and temporal properties of gestures in North American English /r/. *Language and Speech*, *53*, 49–69. DOI: https://doi.org/10.1177/0023830909351209

Christensen, R. H. B. (2019). ordinal—Regression Models for Ordinal Data. R package version 2019.12-10. https://CRAN.R-project.org/package=ordinal

Cohn, A. (2003). Phonological Structure and Phonetic Duration: The Role of the Mora. *Working Papers of the Cornell Phonetics Laboratory*, *15*, 69–100.

Cohn, A., & Lavoie, L. (2003). Superheavy monosyllables in American English. Unpublished manuscript, Cornell University and MIT.

Côté, M. H., & Kharlamov, V. (2011). The impact of experimental tasks on syllabification judgments: A case study of Russian. In C. E. Cairns & E. Raimy (Eds.), *Handbook of the syllable* (pp. 273–294). Leiden: Brill.

Davies, M. (2008). *The Corpus of Contemporary American English (COCA)*. Available online at https://www.english-corpora.org/coca/.

Delattre, P., & Donald, C. F. (1968). A dialect study of American r's by x-ray motion picture. *Linguistics*, *6*, 29–68. DOI: https://doi.org/10.1515/ling.1968.6.44.29

Derwing, B. L., & Eddington, D. (2014). The experimental investigation of syllable structure. *The Mental Lexicon*, *9*(2), 170–195. DOI: https://doi.org/10.1075/ml.9.2.02der

Espy-Wilson, C. Y., Boyce, S. E., Jackson, M., Nayaranan, S., & Alwan, A. (2000). Acoustic modelling of American English /r/. *The Journal of the Acoustical Society of America*, *108*, 343–356. DOI: https://doi.org/10.1121/1.429469

Fant, G. (1960). *Acoustic Theory of Speech Production*. Mouton. The Hague, Netherlands.

Fowler, C., & Saltzman, E. (1993). Coordination and Coarticulation in Speech Production. *Language and Speech, 36*(2–3), 171–195. DOI: https://doi.org/10.1177/002383099303600304

Fricke, S., Szczerbinski, M., Fox-Boyer, A., & Stackhouse, J. (2016). Preschool predictors of early literacy acquisition in German-speaking children. *Reading Research Quarterly, 51*, 29–53. DOI: https://doi.org/10.1002/rrq.116

Geumann, A., Kroos, C., & Tillmann, H. G. (1999). Are there compensatory effects in natural speech? In: *Proceedings of the 14th International Congress of Phonetic Sciences*, 399–402.

Gick, B., & Campbell, F. (2003). Intergestural timing in English /r/. In: Proceedings of the XVth International Congress of Phonetic Sciences. Barcelona: Universitat Autònoma de Barcelona (pp. 1911–1914).

Gick, B., & Goldstein, L. (2002). Relative timing in the three gestures of North American English /r/. *The Journal of the Acoustical Society of America*, *115*, 2481. DOI: https://doi.org/10.1121/1.4778623

Gick, B., Min Kang, A., & Whalen, D. H. (2002). MRI evidence for commonality in the post-oral articulations of English vowels and liquids. *Journal of Phonetics*, *30*(3), 357–371. DOI: https://doi.org/10.1006/jpho.2001.0161

Giles, S., B., & Moll, K. L. (1975). Cinefluorographic study of selected allophones of English /l/. *Phonetica*, *31*, 206–227. DOI: https://doi.org/10.1159/000259670

Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In: *Proceedings of the 8th International Language Resources and Evaluation (LREC ,12)*.

Goldstein, L., & Fowler, C. (2003). Articulatory phonology: A phonology for public language use. In N. O. Schiller & A. S. Meyer (Eds.), *Phonetics and Phonology in Language Comprehension and Production* (pp. 159–207). Mouton de Gruyter. DOI: https://doi.org/10.1515/9783110895094.159

Goldstein, L., Nam, H., Saltzman, E., & Chitoran, I. (2009). Coupled oscillator planning model of speech timing and syllable structure. In C. G. M. Fant, H. Fujisaki & J. Shen (Eds.), *Frontiers in phonetics and speech science* (pp. 239–249).

Göschel, J. (1971). Artikulation und Distribution der sogenannten Liquida r in den europäischen Sprachen. In: *Indogermanische Forschungen*, *76*, 84–126.

Green, A. D. (2001). American English 'r-colored' vowels as complex segments. *Linguistics in Potsdam*, *15*, 70–78.

Hall, T. A. (1993). The phonology of German /R/. *Phonology*, *10*, 83–105. DOI: https://doi.org/10.1017/S0952675700001743

Hammond, M. (1999). The Phonology of English: A prosodic Optimality-Theoretic Approach. Oxford: Oxford University Press.

Hardcastle, W. J., & Barry, W. (1989). Articulatory and perceptual factors in /l/ vocalisation in English. *Journal of the International Phonetics Association*, *15*, 3–17. DOI: https://doi.org/10.1017/S0025100300002930

Harris, J. (1994). English Sound Structure. Oxford: Blackwell.

Hayes, B. (1989). Compensatory lengthening in moraic phonology. *Linguistic Inquiry*, *20*, 253–306.

Heffner, R.-M. S. (1950). General Phonetics. Madison: University of Wisconsin Press.

Honorof, D. N., & Browman, C. P. (1995). The center or edge: How are consonant clusters organised with respect to the vowel? In K. Elenius & P. Branderud (Eds.), *Proceedings of the XIIIth International Congress of Phonetic Sciences*, *3*, 552–555. Stockholm, Sweden: Congress Organisers at KTH and Stockholm University.

Jakobson, R. C., Fant, G. M., & Halle, M. (1952). Preliminaries to speech analysis: The distinctive features and their correlates. The MIT Press, Cambridge, MA.

Kent, R. D., & Moll, K., L. (1972). Tongue Body Articulation during vowel and diphthongs gestures. *Folia Phoniatrica*, *24*, 278–300. DOI: https://doi.org/10.1159/000263574

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*. DOI: https://doi.org/10.18637/jss.v082.i13

Lavoie, L., & Cohn, A. (1999). Sesquisyllables of English: The structure of vowel-liquid syllables. In *Proceedings of the XIVth International Congress of Phonetic Sciences*, 109–112.

Luce, R. D. (1986). Response times: Their role in inferring elementary mental organization. Oxford: OUP.

Marin, S., & Pouplier, M. (2010). Temporal organization of complex onsets and codas in American English: Testing the predictions of a gestural coupling model. *Motor Control*, *14*(3), 380–407. DOI: https://doi.org/10.1123/mcj.14.3.380

Marin, S., & Pouplier, M. (2014). Articulatory synergies in the temporal organization of liquid clusters in Romanian. *Journal of Phonetics*, *42*, 24–36. DOI: https://doi.org/10.1016/j.wocn.2013.11.001

Matisoff, J. (1973). Tonogenesis in Southeast Asia. In Hyman, L. (Ed.) Consonant types and tone. Southern California Occasional Papers in Linguistics No. 1. Los Angeles: USC. 71–96.

Matisoff, J. (1989). The bulging monosyllable, or the mora the merrier: Echo-vowel adverbialization in Lahu. In J. H. C. S. Davidson (Ed.), *South-East Asian Linguistics: Essays in Honor of Eugénie J. A. Henderson. SOAS* (pp. 163–198). University of London.

Mielke, J., et al. (2016). Individual-level contact limits phonological complexity: Evidence from bunched and retroflex /r/. *Language*, *92*, 101–140. DOI: https://doi.org/10.1353/lan.2016.0019

Mooshammer, C., & Schiller, N. O. (1996). Coarticulatory effects on kinematic parameters of rhotics in German. Proceedings of the 1st ESCA Tutorial and Research Workshop on Speech Production Modeling: From Control Strategies to Acoustics & 4th Speech Production Seminar: Models and Data, Autrans, France, May 20–24, 25–28. Grenoble: European Speech Communication Association.

Nam, H., Goldstein, L., & Saltzman, E. (2009). Self-organization of syllable structure: A coupled oscillator model. In M. Pellegrino & Chitoran, C (Eds.), *Approaches to phonological complexity* (pp. 299–328). DOI: https://doi.org/10.1515/9783110223958.297

Nam, H., Goldstein, L., Saltzman, E., & Byrd, D. (2004). TADA: An enhanced, portable task dynamics model in MATLAB. *The Journal of the Acoustical Society of America*, *115*(5,2), 2430. DOI: https://doi.org/10.1121/1.4781490

Narayanan, S., Alwan, A., & Haker, K. (1997). Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part I. The laterals. *The Journal of the Acoustical Society of America, 101*, 1064–1077. DOI: https://doi.org/10.1121/1.418030

Pastätter, M., & Pouplier, M. (2014). The articulatory modelling of German coronal consonants using TADA. *Proceedings of the 10th International Seminar on Speech Production*, Cologne, Germany, 5–8 May 2014.

Popescu, A. (2019). Syllable-count judgments: Relating gestural composition and syllable weight. E. Ronai, L. Stigliano & Y. Sun (Eds.), *Proceedings of the 54th Annual Meeting of the Chicago Linguistics Society* (pp. 425–434).

Pouplier, M. (2012). The gestural approach to syllable structure: Universal, language- and cluster-specific aspects. In W. Fuchs & P. Pape (Eds.), *Speech Planning and Dynamics* (pp. 63–96). Peter Lang.

Proctor, M. (2009). Gestural characterization of a phonological class: The liquids. Doctoral dissertation, Yale University.

Proctor, M., Walker, R., Smith, C., Szalay, T., Louis Goldstein, L., & Narayanan, S. (2019). Articulatory characterization of English liquid-final rimes. *Journal of Phonetics*, *77*, 100921. ISSN 0095-4470. DOI: https://doi.org/10.1016/j.wocn.2019.100921

Recasens, D. (2012). A cross-language acoustic study of initial and final allophones of /l/. *Speech Communication*, *54*, 368–383. DOI: https://doi.org/10.1016/j.specom.2011.10.001

Recasens, D., Pallars, M. D., & Fontdevila, J. (1997). A model of lingual coarticulation based on articulatory constraints. *The Journal of the Acoustical Society of America*, *102*, 544–561. DOI: https://doi.org/10.1121/1.419727

Recasens, D., Pallarès, M. D., & Fontdevila, J. (1998). An electropalatographic and acoustic study of temporal coarticulation for Catalan dark/l/ and German clear/l/. *Phonetica*, *55*(1–2), 53–79. PMID: 9693344. DOI: https://doi.org/10.1159/000028424

Saltzman, E. L., & Munhall, K. M. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, *1*, 333–382. DOI: https://doi.org/10.1207/s15326969eco0104_2

Schiller, N. O., & Mooshammer, C. (1995). The character of /r/-sounds: Articulatory evidence for different reduction processes with special reference to German. In: K. Elenius & P. Branderod (Eds.), *Proceedings of the 13th ICPhS, Stockholm, Sweden, 13–19 August, 1995*, *3*, 452–455. Stockholm: KTH and Stockholm University.

Sproat, R., & Fujimura, O. (1993). Allophonic variation in English /l/ and its implications for phonetic implementation. *Journal of Phonetics*, *21*, 291–311. DOI: https://doi.org/10.1016/S0095-4470(19)31340-3

Stemberger, J. P. (1981). Post-vocalic and syllabic /r/ and /l/ in English. *The Journal of the Acoustical Society of America*, *70*, S40. DOI: https://doi.org/10.1121/1.2018863

Stevens, K. N. (1998). Acoustic Phonetics. The MIT Press, Cambridge, MA/London.

Strycharczuk, P., Derrick, D., & Shaw, J. (2020). Locating de-lateralization in the pathway of sound changes affecting coda /l/. *Laboratory Phonology*, *11*(1), DOI: https://doi.org/10.5334/labphon.236

Tilsen, S. (2014). Selection and coordination of articulatory gestures in temporally constrained production. *Journal of Phonetics*, *44*, 26–46. DOI: https://doi.org/10.1016/j.wocn.2013.12.004

Tilsen, S. (2016). Selection and coordination: The articulatory basis for the emergence of phonological structure. *Journal of Phonetics*, *55*, 53–77. DOI: https://doi.org/10.1016/j.wocn.2015.11.005

Tilsen, S., & Cohn, A. (2016). Shared representations underlie metaphonological judgments and speech motor control. *Laboratory Phonology*, *7*(1), 14. 1–33. DOI: https://doi.org/10.5334/labphon.52

Treiman, R. (1984). On the status of final consonant clusters in English syllables. *Journal of Verbal Learning and Verbal Behaviour*, *23*, 343–356. DOI: https://doi.org/10.1016/S0022-5371(84)90237-8

Turvey, M. T. (1990). Coordination. *American Psychologist*, *45*, 938–953. DOI: https://doi.org/10.1037/0003-066X.45.8.938

Walker, R., & Proctor, M. (2019). The organization and structure of rhotics in American English rhymes. *Phonology*, *36*, 457–485. DOI: https://doi.org/10.1017/S0952675719000228

Whelan, R. (2008). Effective analysis of reaction time data. *The Psychological Record*, *58*(3), 475–482. DOI: https://doi.org/10.1007/BF03395630

Wiese, R. (2003). The unity and variation of (German) /r/. *Zeitschrift für Dialektologie und Linguistik*, *70*(1), 25–43.

Ying, J., Shaw, J. A., Carignan, C., Proctor, M., Derrick, D., & Best, C. T. (2021). Evidence for active control of tongue lateralization in Australian English /l/. *Journal of Phonetics*, *86*. DOI: https://doi.org/10.1016/j.wocn.2021.101039

Yuan, J., & Liberman, M. Y. (2009). Investigating /l/ variation in English through forced alignment. *Proceedings Interspeech 2009* (pp. 2215–2218). DOI: https://doi.org/10.21437/Interspeech.2009-630

Zawadzki, P. A., & David, P. K. (1980). A cineradiographic study of static and dynamic aspects of American English /r/. *Phonetica*, *37*, 253–266. DOI: https://doi.org/10.1159/000259995

Ziegler, J. C., & Goswami, U. (2005). Reading Acquisition, Developmental Dyslexia, and Skilled Reading across Languages: A Psycholinguistic Grain Size Theory. *Psychological Bulletin*, *131*, 3–29. DOI: https://doi.org/10.1037/0033-2909.131.1.3