DISSERTATION / DOCTORAL THESIS

Titel der Dissertation /Title of the Doctoral Thesis

„The role of prosody in language learning, language change and language evolution“

verfasst von / submitted by

Mag. Theresa Matzinger, BSc MSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doctor of Philosophy (PhD)

Wien, 2021 / Vienna 2021

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 794 685 437
Dissertationsgebiet lt. Studienblatt / field of study as it appears on the student record sheet:	Doktoratsstudium NAWI, Bereich Lebenswissenschaften (Dissertationsgebiet: Biologie)
Betreut von / Supervisor:	Univ.-Prof. Tecumseh Fitch, PhD Univ.-Prof. Mag. Dr. Nikolaus Ritt

Acknowledgements

Many people motivated and inspired me during the time of my PhD studies. Thank you!

First, I would like to thank my supervisors Prof. Tecumseh Fitch and Prof. Nikolaus Ritt: Thanks, Niki, for offering me this PhD position, for giving me the freedom to work on the topics that I wanted, for your thought-provoking ideas, and for creating an inspiring and encouraging atmosphere in our team. Thanks, Tecumseh, for your continuous support, your stimulating views and for inviting me to join the DK, where I took part in many exciting events and had fun times at PhD retreats.

Heartful thanks also go to my colleagues, many of whom became close friends of mine: A big thank you goes to my friends and colleagues at the English department, most of all Magdalena, Vanja, Andreas and Klaus, whom it was always a great pleasure to work with and who gave me invaluable advice in many respects. Also, the whole pre-doc community at the English department was never tired of sharing a drink and joking around. Thanks also to Felix, Tudor and the other members of the Fitch lab and CogBio department, who were always open for discussing my ideas and for hanging out together. Thank you, Eva, for making the work on the aesthetics project such a pleasurable and smooth experience. Also, many welcoming and brilliant people from the linguistics, cognitive science and language evolution communities made my conference trips so exciting, inspiring, and fun. Thank you!

Also, I would like to express my gratitude to the team of the BORG II Lessinggasse, who made it possible for me to continue part-time teaching at a high-school level during the time of my PhD and who filled in for me when I couldn’t be at school because of my PhD duties. Thanks for enabling me to also gain professional experience outside the ivory tower of academia.

Finally, I would like to thank my friends, my partner, and my family: Our joint activities contributed to a healthy work-life balance, made me relax and energized me after days and weeks of work. You always managed to cheer me up and lift my spirits, and supported me no matter which direction I went. Thanks for being around and being the awesome people that you are!

Contents

Introduction 4

Chapter 1: Voice modulatory cues to structure across languages and species
Theresa Matzinger & Tecumseh Fitch 28

Chapter 2: The influence of different prosodic cues on word segmentation
Theresa Matzinger, Nikolaus Ritt & Tecumseh Fitch 49

Chapter 3: Aesthetic appeal of prosodic patterns influences their segmentation from a speech stream
Theresa Matzinger, Eva Specker, Nikolaus Ritt & Tecumseh Fitch 78

Chapter 4: Non-native pause patterns closely correspond to those of native speakers at different speech rates
Theresa Matzinger, Nikolaus Ritt & Tecumseh Fitch 121

Chapter 5: Phonotactically probable word shapes represent attractors in the cultural evolution of sound patterns
Theresa Matzinger & Nikolaus Ritt 150

Discussion 178

Abstract 188

Zusammenfassung 190

Introduction

Spoken language is the main acoustic communication system of humans. One of the most fascinating aspects of spoken language is that although its biological bases are shared across humans, characteristics of speech such as stress patterns or speech sounds vary greatly across different dialects or languages (Dryer & Haspelmath, 2013). Also, while humans largely share their phylogeny with other tetrapod species, human spoken language clearly differs from other tetrapods' vocal communication systems (Fitch, 2005; Tallerman & Gibson, 2012). Why is that so?

One of the reasons why speech patterns differ across dialects and languages is that dialects and languages change over time. For example, such changes can be observed on a small scale by listening to children who speak differently than their grandparents or on a larger scale by comparing the speech of present-day English speakers to the speech of their ancestors in the Middle Ages (Fig. 1; Aitchinson, 2001; Bybee, 2015; Chambers & Schilling, 2013; Ritt, 2004). Spoken language in different speech communities may undergo different changes, which creates dialects and may eventually lead to different languages (Croft, 2000; Evans & Levinson, 2009; Keller, 1994; Smith, 2006). Similar pressures may have acted over even longer timespans during the emergence of language in the process of human evolution, and may be responsible for differences in human language and non-human tetrapod vocalizations (Christiansen & Kirby, 2003b; Fitch, 2010). But what exactly are the mechanisms that make languages change the way they do and give them their present-day shapes?

Investigating these and similar questions has a long-standing tradition. When developing his theory of evolution, already Darwin anecdotally highlighted parallels between human languages and animal vocalizations, and pointed towards a common descent of these communication systems (Darwin, 1871). The systematic exploration of language evolution is often taken to have started with the Neogrammarians in the second half of the 19th century (Osthoff & Brugmann, 1878; Paul, 1880), and continues to fascinate linguists, cognitive scientists, and animal communication researchers up to the present day. Especially over the last decades, a vibrant and widely recognized community of researchers (Christiansen & Kirby, 2003b; Dediu & De Boer, 2016; Fitch, 2017; Ravignani et al., 2020) has dedicated itself to the investigation of what makes languages the way they are. Still, the factors that influence language evolution and change are only insufficiently understood (Bybee, 2010; Christiansen

& Chater, 2016; Christiansen & Kirby, 2003a; Hauser et al., 2014). One of the biggest unresolved questions in linguistics and language evolution is how biology and culture interact to give languages their present-day shapes (Kirby, 2017).

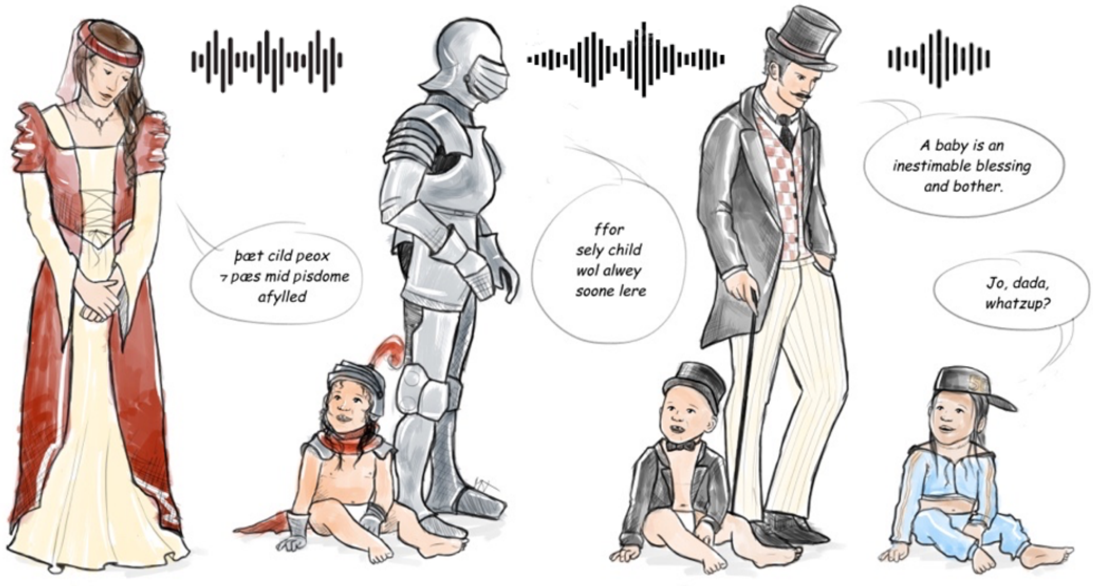


Figure 1. Linguistic change from Old English (Ælfric – Homilies: "the child grew and was filled with wisdom") over Middle English (Geoffrey Chaucer – Canterbury Tales "for an innocent child will always quickly learn") and Modern English (Mark Twain) to Present-Day English. Illustration by Marie-Therese Pekny.

To contribute to this picture, this thesis explores language at the interface of biological and cultural evolution by including research on non-human animal communication systems, human speech production and perception, and actual historical language data. By doing that, it contributes to explaining widely discussed issues in the fields of language evolution and change, such as:

- What are similarities and differences in the vocal production of humans and other tetrapods?
- What are similarities and differences in the production, perception and processing of cross-linguistic and language-specific prosodic patterns?
- What can we conclude from cross-linguistic and cross-species vocal patterns about the underlying physiological pressures and cognitive biases that influence language evolution and change?

- What are the factors that influence how dialects and languages change over time and therefore contribute to the emergence of different languages?
- How do synchronic learning biases change language diachronically?

These questions are closely intertwined, and their broadness and complexity make it impossible for a single thesis to answer any one of them to the extent that they would deserve. The aim of this thesis is therefore to cover selected aspects that contribute to answering these general questions.

Specifically, the thesis focuses on how prosody, which is the acoustic modulation of spoken language (including variations in pitch, duration and speech pauses), contributes to language processing and change by investigating the role of prosody in the identification and the transmission of linguistic structure. The remainder of this chapter will explain how prosody as a means to signal linguistic structure helps to address the interaction of biology and culture during language evolution and change.

Section 1 introduces biological constraints of speech production, perception and processing. Section 2 outlines constraints that act during the cultural transmission of languages and discusses how such production, perception and processing biases can shape linguistic change. Section 3 discusses selected factors that influence how easily linguistic features can be acquired and transmitted from one generation of speakers to the next. It specifically focuses on occurrence frequency, emotions and acoustic salience. Section 4 introduces the prosodic cues investigated in this thesis and explores their role for indicating linguistic patterns and structure. Section 5 specifically highlights the importance of prosody for speech segmentation and discusses why speech segmentation is relevant for recognizing linguistic structure and for investigating linguistic change. Section 6 focuses on the specific research questions addressed and projects conducted within this thesis. Section 7 concludes with on overview of the methodological approaches taken in the thesis.

1. Biological constraints in language transmission

The general goal of communication is to pass on information from one individual to another. However, this process of information transmission from the brain of the speaker to the brain of the listener is limited. To pass information on, speakers first need to encode their thoughts into speech, this speech signal needs to be transmitted acoustically and listeners need to decode the speech signal to extract the informational content (Keller, 1994; Tamariz & Kirby, 2016). This process of encoding, acoustic transmission and decoding of spoken language is constrained by several factors.

First, there are constraints on articulation and auditory perception. The anatomy and physiology of the vocal tract and vocal organs make it impossible for speakers to produce sounds that are, for example, infinitely loud, long, high or low pitched. Similarly, the anatomy and physiology of the hearing organs limit which aspects of acoustic signals listeners can perceive (MacLarnon & Hewitt, 1999; Moon & Lindblom, 2003; Seikel, Drumright, & Hudock, 2021). Secondly, there are constraints on the cognitive processing, learning and memorization of speech signals. For example, speakers and listeners cannot process and memorize infinitely long words or infinitely complex sentences (Baddeley, Gathercole, & Papagno, 1998; Isbilen & Christiansen, 2018; Kurland, 2011; Mishra, 2015).

The above-mentioned constraints are shaped by biological evolution and impose limits on the realization and transmission of spoken language. However, within these biological limits, there is considerable room for variation in the speech signal (Smith et al., 2017). For example, spoken language has many different sounds, and the voice can be acoustically modulated in many different ways (Dryer & Haspelmath, 2013; Hirst & Di Cristo, 1998). Still, although many modulations of the speech signal are biologically possible, some of these speech patterns occur very frequently, whereas others occur only rarely or not at all. For example, across dialects and languages, speakers lengthen syllables (Seifart et al., 2021; Tyler & Cutler, 2009) and decline their pitch at the end of declarative utterances (Hirst & Di Cristo, 1998; Langus, Marchetto, Bion, & Nespor, 2012; Vaissière, 1983), whereas stopping an utterance abruptly is far less common (Berkovits, 1994; Edwards, Beckman, & Fletcher, 1991; Friberg & Sundberg, 1999). This variation can be explained by differences in how easily and efficiently particular speech patterns can be produced, acquired, and transmitted (Croft, 2000; Keller, 1994; Smith et al., 2017). This production, acquisition, and transmission efficiency is the basis for the cultural evolution of languages, which in turn is one of the most important driving factors of language change and the emergence of different languages. The following section explores in more detail what role ease of acquisition and transmission play in the cultural evolution of languages.

2. Cultural transmission of languages as a driving factor of language change

Language stability and language change are driven by the ease of acquisition of speech patterns during the cultural transmission of language from one generation of speakers to the next (Kirby, Cornish, & Smith, 2008; Kirby, Griffiths, & Smith, 2014; Reali & Griffiths, 2009; Tamariz & Kirby, 2016; Tomasello, Kruger, & Ratner, 1993). Individuals learn their speaking behavior by exposure to the speaking behavior of other individuals in their environment, who themselves learned their speaking behavior in the same way (iterated learning; Kirby, Griffiths, & Smith, 2014; Smith, Kirby, & Brighton, 2003). Over many generations of speakers, small biases in production, processing or perception may amplify and lead to linguistic stability or long-term linguistic change.

Speakers may – either completely or partially accidentally (e.g. due to limited experience, mispronunciations or as short-term reactions to interrupting noise in the environment) – produce novel linguistic variants. Also, listeners may accidentally perceive linguistic items erroneously and reproduce them inaccurately in their own speech production, which creates novel variants (Foulkes & Vihman, 2015; Lass, 1997; Ritt, 2004). If these novel variants are less easily acquired than existing variants, they will not spread, and the existing variants will stabilize. However, by chance, in a particular communicative situation or environment, the novel and accidentally produced linguistic variants may be more easily acquired than their progenitor variants, for example because they are particularly salient, memorable, aesthetic (e.g. Kirby et al., 2008; also see section 3 and references therein for more detailed explanations) or socially desirable or prestigious (Labov, 1963, 2001; Lev-Ari & Peperkamp, 2014; Roberts & Fedzechkina, 2018). In this case, learners may adopt and continue to preferentially use these novel linguistic features, while at the same time neglecting old competing variants. This leads to higher occurrence frequencies of the novel patterns and makes the novel patterns more prominent and reliable inputs for future speaker generations (Croft, 2006; Lass, 2003; Zehentner, 2019). Via this mechanism, which is similar to natural selection and driven by the competition of linguistic variants in an energetically constrained environment, novel linguistic patterns can emerge and spread (Keller, 1994; Lass, 2003; Mesoudi, 2007, 2011; Rosenbach, 2008), which can eventually lead to novel dialects and languages.

An important issue in the investigation of the cultural evolution of languages is the question which factors determine if novel linguistic variants are easily recognizable, acquirable, and memorable so that they can spread and stabilize successfully. This will be discussed in the next section.

3. Factors that make spoken language easily acquirable and transmissible

In general, speech patterns can be acquired and transmitted easily when they have a salient, characteristic, and unambiguous structure. Language learners and users need to identify linguistic structures in the signals they are exposed to, associate these structures with their meanings and re-produce these structures in their own speech. The more reliable and unambiguous these structures are, the more effective language learning and use will be (Kirby et al., 2008; Tamariz & Kirby, 2016). There are several factors that make speech patterns reliable and useful cues to structure, help learners to identify these patterns in the speech signal and therefore make language more easily acquirable and transmissible. A subset of these factors, namely those that are most relevant for this thesis, will be discussed in more detail in this section.

First, the occurrence frequency of linguistic structures is an important determiner of how characteristic they are and how easily they can be recognized, acquired and processed (Bybee, 2007; Divjak, 2019). In general, linguistic patterns that are more frequent and thus more probable to occur in the linguistic environment that language learners and users are exposed to, are identified more easily (Frost, Monaghan, & Christiansen, 2019; Kelley & Tucker, 2017), acquired and memorized more easily (Diessel, 2007; Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012; Storkel, 2001), and produced more accurately (Goldrick & Larson, 2008) than less frequent and less probable forms. This is closely linked to the concept of priming, which describes the facilitated processing of utterances with similar linguistic patterns and which leads to the tendency of speakers to subconsciously repeat linguistic features that they have encountered before (Pickering & Branigan, 1999). Processing frequent and predictable linguistic patterns is therefore assumed to be easier than processing rare and unpredictable linguistic patterns, and occurrence frequencies may for this reason also affect linguistic change (Diessel, 2007; Hall, Hume, Jaeger, & Wedel, 2018; Haselow, 2018; Jäger & Rosenbach, 2008).

Secondly, the salience and in turn the processing fluency of linguistic features may be influenced by the affect and emotions that the linguistic features evoke (Esposito, 2020; Forster, 2020; Forster, Leder, & Ansorge, 2013; Lüdtke, 2015; Reber, Schwarz, & Winkielman, 2004).

Emotions are mostly associated with the semantic content of utterances (Foolen, 2015; Sereno, Scott, Yao, Thaden, & O'Donnell, 2015; Warriner, Kuperman, & Brysbaert, 2013), but can also be evoked by speech prosody, such as prosody conveying happiness, sadness or fear (Filippi et al., 2016; Petrone, Carbone, & Champagne-Lavau, 2016; Quam & Swingley, 2012). Linguistic features that cause negative arousal or are highly pleasant may be more striking and may therefore be processed and acquired more easily than linguistic features with average arousal or valence values (Ferré, 2003; Larsen, Mercer, Balota, & Strube, 2008; Warriner et al., 2013). For example, in lexical decision tasks, participants usually respond faster to positive and negative words than to neutral words (Kousta, Vinson, & Vigliocco, 2009; Scott, O'Donnell, & Sereno, 2012; but also see the discussion in Kuperman, Estes, Brysbaert, & Warriner, 2014). Also, highly pleasant words are acquired earlier by first language learners than less pleasant ones (Warriner et al., 2013). Therefore, affective and emotional values of linguistic features may influence the ease of language processing and can in turn also have an effect on linguistic change (Brown, 2017).

Finally, certain linguistic patterns can be more easily perceived and processed than others because of their intrinsic acoustic salience (Boswijk & Coler, 2020; Dziubalska-Kołaczyk, 2019; Rácz, 2013; Zarcone, van Schijndel, Vogels, & Demberg, 2016). For example, pauses may be more effective in structuring spoken language than other prosodic cues because silence is more easily recognizable than variations in pitch or duration. Also, producing pauses is easy compared to the articulation of vocal elements varying in their duration, pitch and vocal quality (Fletcher, 2010). Although it is hard to make generalizations about which specific linguistic features are most salient and easiest producible, it is likely that such differences in the inherent salience of linguistic features can influence linguistic change.

All of these factors help language learners and users to recognize linguistic structures and identify links between these structures and associated meanings. But which acoustic features indicate linguistic structure in the speech signal in the first place? One prominent means for signaling linguistic structure is prosody, which will be explored in the following section.

4. Prosody as a means to structure language

The term prosody is used to subsume all suprasegmental voice modulatory parameters of spoken language such as variations in duration, fundamental frequency ('pitch') or intensity.

Also, prosody includes speech pauses. These voice modulatory cues contribute to the production and perception of stress, speech rhythm, tempo and intonation patterns (Fletcher, 2010; Gussenhoven & Chen, 2020). Historically, the term prosody has excluded voice modulation on the phoneme level. However, when dealing with phenomena such as word stress or rhythm, it is often impossible to only focus on suprasegmental features without regarding segmental features such as inherent vowel length. Thus, henceforth, prosody will be conceptualized broadly as "the abstract hierarchical phonological structure(s) of an utterance, and prominence relations within that structure" (Fletcher, 2010, p. 523).

Prosodic modulations create patterns in the speech signal, which are often indications of underlying meaning-associated structures such as morphological, syntactic or semantic structures (Fletcher, 2010; Shattuck-Hufnagel & Turk, 1996). Some prosodic structures occur across languages, whereas others are language-specific. For example, the cross-linguistically consistent lengthening of elements at the end of phrases (Seifart et al., 2021; Tyler & Cutler, 2009) and a pitch decrease at the end of declarative sentences can serve as boundary signals for listeners (Hirst & Di Cristo, 1998; Langus et al., 2012; Vaissière, 1983). On the other hand, language-specific prosodic patterns such as word-stress patterns can help to identify and process words (Dryer & Haspelmath, 2013; Ordin, Polyanskaya, Laka, & Nespor, 2017; Slowiaczek, 1990; Tyler & Cutler, 2009). Also, non-human tetrapods are able to modulate the prosodic features of their vocalizations, which indicates structure in their signals (Filippi, Hoeschele, Spierings, & Bowling, 2019; Garcia & Favaro, 2017; Kershenbaum et al., 2016). The interaction of cross-linguistically consistent and language-specific prosodic cues for signaling linguistic structure, and the occurrence of similar cues in non-human tetrapod vocalizations make prosody a very interesting target for investigating the interaction of biology and culture in language evolution and change.

Prosodic cues are not only linked to cognitive abilities such as pattern recognition and memory skills, but also to physiological properties of the vocal apparatus. Variations of pitch are produced by the vibration of the vocal folds, and variations of duration and intensity by the duration and force of exhalation of air from the lungs (Fitch, 2010; Pisanski, Cartei, McGettigan, Raine, & Reby, 2016; Taylor & Reby, 2010). Pauses are influenced by the speakers' breathing range (Fletcher, 2010; MacLarnon & Hewitt, 1999). Because prosody is closely linked to physiology and cognition, prosodic features may be subject to biological and cultural evolutionary pressures. Biological evolution may lead to changes in the vocal apparatus that may enable speakers to produce novel prosodic modulations, whereas cultural biases may favor the use of particular prosodic modulations over others. Thus, both biological and cultural pressures can lead to changing prosodic patterns over time (Brown, 2017; Gussenhoven & Chen, 2020; Lahiri, Riad, & Jacobs, 1999; Smith, 2011; Tallerman & Gibson, 2012). Also, since prosodic cues have many different acoustic correlates and production mechanisms (Dogil & Williams, 1999; Gordon & Roettger, 2017; Gussenhoven & Chen, 2020), different cues may all have distinctive and separable effects on speech processing, may develop differently over time, and are therefore worth being investigated separately. Because prosody is at the interface of physiology and cognition, the investigation of prosodic cues is well-suited to shed light on the role of biology and culture in language evolution and change.

The next section will explore how exactly prosodic cues can help language learners and users to recognize, memorize and encode linguistic structure and how this can influence the transmission of spoken language during cultural evolution.

5. The role of prosody in speech segmentation

One crucial problem in language processing and in perceiving linguistic structure is to efficiently segment continuous speech into words. This problem, which is commonly known as the speech segmentation problem, is a challenge that is most acute for infant first language learners but is equally faced by second language learners and adults listening to a foreign language (Cutler, 1990; Endress & Hauser, 2010; Erickson & Thiessen, 2015; Johnson & Jusczyk, 2001). Listeners need to identify boundaries between words, phrases or sentences to be able to extract their semantic content (Singh, Reznick, & Xuehua, 2012; Yurovsky, Yu, & Smith, 2012).

Prosody offers important cues that help language learners and language users to solve the speech segmentation problem and to indicate linguistic structure. Prosodic cues can facilitate speech segmentation in several different ways. On the one hand, speakers can indicate boundaries directly, for example via pauses, durational modifications or pitch modifications at boundaries (Cutler, Dahan, & Van Donselaar, 1997; Hawthorne & Gerken, 2014; Holzgrefe-Lang et al., 2016; Petrone et al., 2017; Saffran, Newport, & Aslin, 1996; Wellmann, Holzgrefe, Truckenbrodt, Wartenburger, & Höhle, 2012). These signals are assumed to be extracted by listeners relatively directly and to require only a limited amount of cognitive processing (Johnson & Jusczyk, 2001; Johnson & Seidl, 2009; Thiessen & Saffran, 2003).

On the other hand, prosodic cues can give words prototypical shapes, for example by creating word stress or tone patterns (Dryer & Haspelmath, 2013; Gordon & Roettger, 2017). Listeners may use their knowledge about the typical word stress patterns of the dialects and languages they hear to infer which syllables belong together to make up a word (Ordin & Nespor, 2013; Ordin et al., 2017; Tyler & Cutler, 2009). On a more fine-grained level, prototypical prosodic patterns may even serve to indicate different word classes or morphological patterns (cf. Baumann, Prömer, & Ritt, 2019; Dziubalska-Kołaczyk, 2019; Post, Marslen-Wilson, Randall, & Tyler, 2008). Using prototypical prosodic patterns as an indicator for linguistic structure is assumed to require more cognitive effort than simply using immediate prosodic cues as boundary indicators because it requires tracking, monitoring and memorizing the word stress patterns of the languages that listeners are exposed to, and these patterns are often language-specific.

Additionally, specific prosodic patterns may be particularly salient, aesthetically appealing, or create arousal or valence, which may bias listeners towards perceiving these patterns as units in the speech stream. Thus, overall, prosodic cues can facilitate speech segmentation and language acquisition in many different ways.

6. Overview of the thesis chapters and specific research questions addressed

Since prosody facilitates speech segmentation and signals linguistic structure, it also influences how easily particular speech patterns are acquired and transmitted, and in turn how languages change diachronically. To explore these relationships in more detail, the core question of this thesis is “How does prosody contribute to signaling linguistic structure, and in turn to language processing and change?”. Addressing this question sheds light on how physiological constraints of articulating, processing and perceiving particular prosodic patterns interact with factors that are important in the cultural transmission of languages, such as occurrence frequencies, aesthetic appeal or salience.

Constraints on the production, perception and processing of prosodic patterns can be tested experimentally and against historical language data and can be put in perspective in comparison to non-human tetrapod vocalizations. This thesis combines these approaches in five different projects.

The first project (**Chapter 1**), accepted for publication in *Philosophical Transactions of the Royal Society B*, compares prosodic cues to structure across languages and species in a review article. If specific prosodic patterns occur cross-linguistically and across species, this suggests that physiological or basic cognitive processing constraints may be responsible for these patterns. These constraints may be evolutionarily old and may be shared across dialects, languages and species because of a common ancestry or shared environmental pressures. In contrast, prosodic cues that differ across dialects, languages or species may be mainly influenced by different environmental pressures and cultural evolutionary processes that amplify small biases in different directions (Christiansen, Collins, & Edelman, 2009; Croft, 2010; Evans & Levinson, 2009; Fishbein, Fritz, Idsardi, & Wilkinson, 2019; Fitch, 2005, 2010; Ravignani et al., 2019).

The second project (**Chapter 2**), published in *Frontiers in Psychology* (Matzinger, Ritt, & Fitch, 2021), presents an experimental investigation of how effective different prosodic cues such as pauses, durational cues and pitch cues are for perceiving words in a novel language, using an artificial language learning paradigm. Participants segmented words from a continuous stretch of speech of an artificial language, and we measured how well they perceived words when the speech stream was prosodically modified. The project also tests if prosodic cues realized at different positions in a word have a different influence on which words are perceived. This gives insights on the relative importance of different prosodic cues for speech segmentation and on whether cross-linguistically consistent or language-specific prosodic cues are more helpful for learners to acquire a novel language.

The third project (**Chapter 3**), currently submitted and in review at the *Journal of Language Evolution*, contributes to answering the question of whether aesthetic perception of prosodic features may determine how well words are acquired and transmitted. This project experimentally tests if the aesthetic value of different prosodic patterns differs and if cross-linguistically consistent or language-specific prosodic patterns are perceived as more aesthetic. Participants listened to a set of prosodically modified words from an artificial language and rated these words on their aesthetic appeal. By using similar stimuli as in Chapter 2, the project links the aesthetic appeal of various prosodic patterns to their effectiveness for word perception investigated in Chapter 2.

The fourth project (**Chapter 4**), published in *PLoS ONE* (Matzinger, Ritt, & Fitch, 2020), investigates how well highly acoustically salient cues such as speech pauses are learned and realized when speaking a second language and at which positions within a text they are placed to signal linguistic structure. First and second language speakers of English read a standardized English text at three different speech rates (fast, casual and slow), and we measured the characteristics of their speech pauses, such as pause duration, number or position in the text. Similarities and differences in the pause patterns of first and second language speakers can reveal whether pauses are mainly determined by language-specific cultural pressures or by more general physiological or cognitive capacities such as breathing range, attention or memory.

The fifth project (**Chapter 5**), currently submitted and in review at *Cognitive Linguistics*, is a corpus study that uses one particular Middle English sound change, known as Open Syllable Lengthening, as a model to investigate if the occurrence frequency of particular prosodic patterns can influence how the prosodic structure of a language changes diachronically. Further, it tests if language users are sensitive to correspondences between prosodic patterns and linguistic structures not only on the level of words but also on more fine-grained levels such as the morphological level.

Thus, taken together, the five projects a) investigate several different factors, i.e. acoustic salience, aesthetic appeal and occurrence frequency, that may determine how well different prosodic patterns are acquired and transmitted to future speaker generations, b) test if these transmission factors act on both the production and perception of prosodic features, c) test if these factors can account for actual diachronic changes of prosodic patterns and d) discuss similarities and differences of prosodic patterns in human languages and non-human tetrapod vocalizations. Overall, these projects contribute to clarifying the role of prosody during language transmission and change, and thus address the interaction of biological and cultural factors in language evolution and change.

7. Methodological choice – an interdisciplinary dissertation project

This thesis pursues an interdisciplinary approach and uses a mix of methods to address its research questions. The first project of the thesis is a literature review on voice modulatory cues in human and non-human tetrapod vocalizations (**Chapter 1**). It is followed by the empirical investigation of a) natural and artificial languages, b) synchronic and diachronic data, and c) production and perception data: in the second project (**Chapter 2**), I conducted an artificial language learning experiment on the perception of different prosodic patterns. Artificial language learning is widely established and highly informative in the study of language learning, language transmission and language evolution (Culbertson & Schuler, 2019; Kirby et al., 2008; Saffran, Aslin, & Newport, 1996; Smith et al., 2017). In the third project (**Chapter 3**), I used methods from empirical aesthetics, which have previously mostly been used for visual stimuli (Leder, Ring, & Dressler, 2013; Skov & Nadal, 2020), and innovatively apply them to acoustic stimuli. This involved the collection of valence ratings of prosodic patterns in an artificial language. In the fourth project (**Chapter 4**), I collected high quality speech recordings of English native and non-native speakers, inspired by methods from phonetics and second language research (Munro & Derwing, 1995; Piske, MacKay, & Flege, 2001; Robinson & Ellis, 2008; Styler, 2017) and analyzed the pauses that speakers made. In the fifth project (**Chapter 5**), I investigated occurrence frequency effects in diachronic corpus data (Diessel, 2007; Honeybone & Salmons, 2015; Stefanowitsch, 2020). Using long-term historical language data is ideal for testing if experimental findings on the production and perception of prosodic features in present-day or artificial languages can also explain the historical development of actual languages. This study also addresses how to gain insights on prosodic features, which are normally spoken, from written corpus data.

Thus, in summary, I collected speech perception data, valence ratings, speech recordings and diachronic corpus data, and the thesis combines methods from the fields of psycholinguistics, empirical aesthetics, second language research, and diachronic corpus linguistics. Additionally, it discusses findings from animal communication research. Theoretically, my dissertation is based within a cognitive framework and is deeply rooted in evolutionary thinking, inspired by biological evolution (Aldrich et al., 2008; Croft, 2002; Fitch, 2010; Pagel, 2017; Ritt, 2004; Smith, 2006). This set of methods and theoretical frameworks addresses the role of prosody in language evolution and change from different perspectives and therefore contributes to a more complete picture of the role of prosody in language evolution and change.

8. References

Aitchinson, J. (2001). *Language change: progress or decay?* Cambridge: Cambridge University Press.

Aldrich, H. E., Hodgson, G. M., Hull, D. L., Knudsen, T., Mokyr, J., & Vanberg, V. J. (2008). In defence of generalized Darwinism. *Journal of Evolutionary Economics*, *18*(5), 577–596. https://doi.org/10.1007/s00191-008-0110-z

Baddeley, A. D., Gathercole, S. E., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review*, *105*(1), 158–173. https://doi.org/10.4324/9781315111261

Baumann, A., Prömer, C., & Ritt, N. (2019). Word form shapes are selected to be morphotactically indicative. *Folia Linguistica*, *40*(1), 129–151. https://doi.org/10.1515/flih-2019-0007

Berkovits, R. (1994). Durational effects in final lengthening, gapping, and contrastive stress. *Language and Speech*, *37*(3), 237–250. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7861912

Boswijk, V., & Coler, M. (2020). What is Salience? *Open Linguistics*, *6*(1), 713–722. https://doi.org/10.1515/opli-2020-0042

Brown, S. (2017). A joint prosodic origin of language and music. *Frontiers in Psychology*, *8*, 1–20. https://doi.org/10.3389/fpsyg.2017.01894

Bybee, J. (2007). *Frequency of use and the organization of language*. Oxford: Oxford University Press.

Bybee, J. (2010). *Language, usage and cognition*. Cambridge: Cambridge University Press.

Bybee, J. (2015). *Language change*. Cambridge: Cambridge University Press.

Chambers, J. K., & Schilling, N. (Eds.). (2013). *The Handbook of Language Variation and Change*. *The Handbook of Language Variation and Change* (2nd ed.). Chichester: Wiley-Blackwell. https://doi.org/10.1111/b.9781405116923.2003.00039.x

Christiansen, M. H., & Chater, N. (2016). *Creating language - integrating evolution, acquisition, and processing*. Cambrdige, Mass.: MIT Press.

Christiansen, M. H., Collins, C., & Edelman, S. (Eds.). (2009). *Language universals*. Oxford: Oxford University Press.

Christiansen, M. H., & Kirby, S. (2003a). Language evolution: Consensus and controversies. *Trends in Cognitive Sciences*, *7*(7), 300–307. https://doi.org/10.1016/S1364-6613(03)00136-0

Christiansen, M. H., & Kirby, S. (Eds.). (2003b). *Language Evolution*. Oxford: Oxford University Press.

Croft, W. (2000). *Explaining Language Change - An Evolutionary Approach*. Hartow: Pearson Education Limited.

Croft, W. (2002). The Darwinization of Linguistics. *Selection*, *3*(1), 75–91. https://doi.org/10.1556/Select.3.2002.1.7

Croft, W. (2006). Evolutionary models and functional-typological theories of language change. In A. Van Kemenade & B. Los (Eds.), *The handbook of English linguistics* (pp. 68–92). Malden, MA: Blackwell.

Croft, W. (2010). Relativity, linguistic variation and language universals. *CogniTextes*, *4*(4). https://doi.org/10.4000/cognitextes.303

Culbertson, J., & Schuler, K. (2019). Artificial Language Learning in Children. *Annual Review of Linguistics*, *5*, 353–373. https://doi.org/10.1146/annurev-linguistics-011718-012329

Cutler, A. (1990). Exploiting Prosodic Probabilities in Speech Segmentation. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing* (pp. 105–121). Cambridge, MA: MIT Press.

Cutler, A., Dahan, D., & Van Donselaar, W. (1997). Prosody in the Comprehension of Spoken Language : A Literature Review. *Language and Speech*, *40*(2), 141–202.

Darwin, C. (1871). *The descent of man and selection in relation to sex*. London: John Murray.

Dediu, D., & De Boer, B. (2016). Language evolution needs its own journal. *Journal of Language Evolution*, *1*(1), 1–6. https://doi.org/10.1093/jole/lzv001

Diessel, H. (2007). Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology*, *25*, 108–127. https://doi.org/10.1016/j.newideapsych.2007.02.002

Divjak, D. (2019). *Frequency in Language - Memory, Attention and Learning*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316084410

Dogil, G., & Williams, B. (1999). The phonetic manifestation of word stress. In H. van der Hulst (Ed.), *Word prosodic systems in the languages of Europe* (pp. 273–310). Berlin: Mouton de Gruyter.

Dryer, M. S., & Haspelmath, M. (2013). The World Atlas of Language Structures Online. Retrieved from https://wals.info

Dziubalska-Kołaczyk, K. (2019). On the structure, survival and change of consonant clusters. *Folia Linguistica*, *40*(1), 107–127. https://doi.org/10.1515/flih-2019-0006

Edwards, J., Beckman, M. E., & Fletcher, J. (1991). The articulatory kinematics of final lengthening. *The Journal of the Acoustical Society of America*, *89*(1), 369–382. https://doi.org/10.1121/1.400674

Endress, A. D., & Hauser, M. D. (2010). Word segmentation with universal prosodic cues. *Cognitive Psychology*, *61*(2), 177–199. https://doi.org/10.1016/j.cogpsych.2010.05.001

Erickson, L. C., & Thiessen, E. D. (2015). Statistical learning of language: Theory, validity, and predictions of a statistical learning account of language acquisition. *Developmental Review*, *37*, 66–108. https://doi.org/10.1016/j.dr.2015.05.002

Esposito, L. (2020). Linking gender, sexuality, and affect: The linguistic and social patterning of phrase-final posttonic lengthening. *Language Variation and Change*, *32*(2), 191–216. https://doi.org/10.1017/S0954394520000095

Evans, N., & Levinson, S. C. (2009). The myth of language universals : Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, (32), 429–492.

Ferré, P. (2003). Effects of level of processing on memory for affectively valenced words. *Cognition & Emotion*, *17*(6), 859–880. https://doi.org/10.1080/02699930244000200

Filippi, P., Hoeschele, M., Spierings, M., & Bowling, D. L. (2019). Temporal modulation in speech, music, and animal vocal communication: evidence of conserved function. *Annals of the New York Academy of Sciences*, *1453*(1), 99–113. https://doi.org/10.1111/nyas.14228

Filippi, P., Ocklenburg, S., Bowling, D. L., Heege, L., Güntürkün, O., Newen, A., & Boer, B. de. (2016). More than words (and faces): evidence for a Stroop effect of prosody in emotion word processing. *Cognition and Emotion*, *0*(0), 1–13. https://doi.org/10.1080/02699931.2016.1177489

Fishbein, A. R., Fritz, J. B., Idsardi, W. J., & Wilkinson, G. S. (2019). What can animal communication teach us about human language? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(20190042), 1–4. https://doi.org/10.1098/rstb.2019.0042

Fitch, W. T. (2005). The Evolution of Language: A Comparative Review. *Biology and Philosophy*, *20*(2–3), 193–203. https://doi.org/10.1007/s10539-005-5597-1

Fitch, W. T. (2010). *The Evolution of Language*. Cambridge University Press.

Fitch, W. T. (2017). Empirical approaches to the study of language evolution. *Psychonomic Bulletin and Review*, *24*(1), 3–33. https://doi.org/10.3758/s13423-017-1236-5

Fletcher, J. (2010). The Prosody of Speech : Timing and Rhythm. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The handbook of phonetic sciences* (2nd ed., pp. 523–

602). Hoboken: Wiley-Blackwell.

Foolen, A. (2015). Word valence and its effects. In U. M. Lüdtke (Ed.), *Emotion in language: theory - research - application*. Amsterdam: John Benjamins.

Forster, M. (2020). Processing fluency. In M. Nadal & O. Vartanian (Eds.), *The Oxford Handbook of Empirical Aesthetics*. Oxford: Oxford University Press.

Forster, M., Leder, H., & Ansorge, U. (2013). It felt fluent, and I liked it: subjective feeling of fluency rather than objective fluency determines liking. *Emotion*, *13*(2), 280–289. https://doi.org/10.1037/a0030115

Foulkes, P., & Vihman, M. (2015). First Language Acquisition and Phonological Change. In P. Honeybone & J. Salmons (Eds.), *Oxford Handbooks Online* (pp. 1–32). Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199232819.013.001

Friberg, A., & Sundberg, J. (1999). Does music performance allude to locomotion ? A model of final ritardandi derived from measurements of stopping runners. *The Journal of the Acoustical Society of America*, *105*(3), 1469–1484. https://doi.org/10.1121/1.426687

Frost, R. L. A., Monaghan, P., & Christiansen, M. H. (2019). Mark My Words: High Frequency Marker Words Impact Early Stages of Language Learning. *Journal of Experimental Psychology: Learning Memory and Cognition*, *45*(10), 1883–1898. https://doi.org/10.1037/xlm0000683

Garcia, M., & Favaro, L. (2017). Animal vocal communication: Function, structures, and production mechanisms. *Current Zoology*, *63*(4), 417–419. https://doi.org/10.1093/cz/zox040

Goldrick, M., & Larson, M. (2008). Phonotactic probability influences speech production. *Cognition*, *107*(3), 1155–1164. https://doi.org/10.1016/j.cognition.2007.11.009

Gordon, M., & Roettger, T. (2017). Acoustic correlates of word stress: A cross-linguistic survey. *Linguistics Vanguard*, *3*(1), 1–11. https://doi.org/10.1515/lingvan-2017-0007

Gussenhoven, C., & Chen, A. (Eds.). (2020). *The Oxford handbook of language prosody*. Oxford: Oxford University Press.

Hall, K. C., Hume, E., Jaeger, T. F., & Wedel, A. (2018). The role of predictability in shaping phonological patterns. *Linguistics Vanguard*, *20170027*, 1–15. https://doi.org/10.1515/lingvan-2017-0027

Haselow, A. (2018). Language change from a psycholinguistic perspective: The long-term effects of frequency on language processing. *Language Sciences*, *68*, 56–77. https://doi.org/10.1016/j.langsci.2017.12.006

Hauser, M. D., Yang, C., Berwick, R. C., Tattersall, I., Ryan, M. J., Watumull, J., …

Lewontin, R. C. (2014). The mystery of language evolution. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00401

Hawthorne, K., & Gerken, L. (2014). From pauses to clauses: prosody facilitates learning of syntactic constituency. *Cognition*, *133*(2), 455–459. https://doi.org/10.1126/science.1249749.Ribosome

Hirst, D., & Di Cristo, A. (Eds.). (1998). *Intonation systems: a survey of twenty languages*. Cambridge: Cambridge University Press.

Holzgrefe-Lang, J., Wellmann, C., Petrone, C., Räling, R., Truckenbrodt, H., Höhle, B., & Wartenburger, I. (2016). How pitch change and final lengthening cue boundary perception in German : converging evidence from ERPs and prosodic judgements. *Language, Cognition and Neuroscience*, *31*(7), 904–920. https://doi.org/10.1080/23273798.2016.1157195

Honeybone, P., & Salmons, J. (Eds.). (2015). *The Oxford Handbook of Historical Phonology*. Oxford: Oxford University Press.

Isbilen, E. S., & Christiansen, M. H. (2018). Chunk-Based Memory Constraints on the Cultural Evolution of Language. *Topics in Cognitive Science*, *12*, 713–726. https://doi.org/10.1111/tops.12376

Jäger, G., & Rosenbach, A. (2008). Priming and unidirectional language change. *Theoretical Linguistics*, *34*(2), 85–113.

Johnson, E. K., & Jusczyk, P. W. (2001). Word Segmentation by 8-Month-Olds : When Speech Cues Count More Than Statistics. *Journal of Memory and Language*, *44*, 548–567. https://doi.org/10.1006/jmla.2000.2755

Johnson, E. K., & Seidl, A. H. (2009). At 11 months, prosody still outranks statistics. *Developmental Science*, *12*(1), 131–141. https://doi.org/10.1111/j.1467-7687.2008.00740.x

Keller, R. (1994). *On language change - the invisible hand in language*. New York: Routledge. https://doi.org/10.4324/9780203993286

Kelley, M. C., & Tucker, B. V. (2017). The effects of phonotactic probability on auditory recognition of pseudo-words. *The Journal of the Acoustical Society of America*, *141*(5), 4038. https://doi.org/10.1121/1.4989319

Kershenbaum, A., Blumstein, D. T., Roch, M. A., Akçay, Ç., Bee, M. A., Bohn, K., … Zamora-Gutierrez, V. (2016). Acoustic sequences in non-human animals: a tutorial review and prospectus. *Biological Reviews of the Cambridge Philosophical Society*, *91*(1), 13–52. https://doi.org/10.1111/brv.12160.Acoustic

Kirby, S. (2017). Culture and biology in the origins of linguistic structure. *Psychological Bulletin*, *24*, 118–137. https://doi.org/10.3758/s13423-016-1166-7

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: an experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(31), 10681–10686. https://doi.org/10.1073/pnas.0707835105

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, *28*, 108–114. https://doi.org/10.1016/j.conb.2014.07.014

Kousta, S. T., Vinson, D. P., & Vigliocco, G. (2009). Emotion words, regardless of polarity, have a processing advantage over neutral words. *Cognition*, *112*(3), 473–481. https://doi.org/10.1016/j.cognition.2009.06.007

Kuperman, V., Estes, Z., Brysbaert, M., & Warriner, A. B. (2014). Emotion and language: valence and arousal affect word recognition. *Journal of Experimental Psychology. General*, *143*(3), 1065–1081. https://doi.org/10.1037/a0035669.Emotion

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990. https://doi.org/10.3758/s13428-012-0210-4

Kurland, J. (2011). The Role That Attention Plays in Language Processing. *Perspectives on Neurophysiology and Neurogenic Speech and Language Disorders*, *21*(2), 47–54. https://doi.org/10.1044/nnsld21.2.47

Labov, W. (1963). The Social Motivation of a Sound Change. *Word*, *19*(3), 273–309. https://doi.org/10.1080/00437956.1963.11659799

Labov, W. (2001). *Principles of linguistic change: social factors*. Malden, MA: Blackwell.

Lahiri, A., Riad, T., & Jacobs, H. (1999). Diachronic prosody. In H. Van der Hulst (Ed.), *Word prosodic systems in the languages of Europe* (pp. 335–424). Berlin: Mouton de Gruyter.

Langus, A., Marchetto, E., Bion, R. A. H., & Nespor, M. (2012). Can prosody be used to discover hierarchical structure in continuous speech? *Journal of Memory and Language*, *66*(1), 285–306. https://doi.org/10.1016/j.jml.2011.09.004

Larsen, R. J., Mercer, K. A., Balota, D. A., & Strube, M. J. (2008). Not All Negative Words Slow Down Lexical Decision and Naming Speed: Importance of Word Arousal. *Emotion*, *8*(4), 445–452. https://doi.org/10.1037/1528-3542.8.4.445

Lass, R. (1997). *Historical linguistics and language change*. Cambridge: Cambridge

University Press.

Lass, R. (2003). Genetic metaphor in historical linguistics. *Alternation*, *10*(1), 47–62.

Leder, H., Ring, A., & Dressler, S. G. (2013). See me, feel me! Aesthetic evaluations of art portraits. *Psychology of Aesthetics, Creativity, and the Arts*, *7*(4), 358–369. https://doi.org/10.1037/a0033311

Lev-Ari, S., & Peperkamp, S. (2014). An experimental study of the role of social factors in language change: The case of loanword adaptations. *Laboratory Phonology*, *5*(3), 379–401. https://doi.org/10.1515/lp-2014-0013

Lüdtke, U. M. (Ed.). (2015). *Emotion in language: theory - research - application*. Amsterdam: John Benjamins.

MacLarnon, A. M., & Hewitt, G. P. (1999). The evolution of human speech: The role of enhanced breathing control. *American Journal of Physical Anthropology*, *109*(3), 341–363. https://doi.org/10.1002/(SICI)1096-8644(199907)109:3<341::AID-AJPA5>3.0.CO;2-2

Matzinger, T., Ritt, N., & Fitch, W. T. (2020). Non-native speaker pause patterns closely correspond to those of native speakers at different speech rates. *PLoS ONE*, *15*(4), 1–20. https://doi.org/10.1371/journal.pone.0230710

Matzinger, T., Ritt, N., & Fitch, W. T. (2021). The Influence of Different Prosodic Cues on Word Segmentation. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.622042

Mesoudi, A. (2007). Biological and Cultural Evolution: Similar but Different. *Biological Theory*, *2*(2), 119–123. https://doi.org/10.1162/biot.2007.2.2.119

Mesoudi, A. (2011). *Cultural evolution - How Darwinian theory can explain human culture & synthesize the social sciences*. Chicago: The University of Chicago Press.

Mishra, R. K. (2015). *Interaction between attention and language systems in humans: A cognitive science perspective*. New Delhi: Springer. https://doi.org/10.1007/978-81-322-2592-8

Moon, S.-J., & Lindblom, B. (2003). Two experiments on oxygen consumption during speech production: vocal effort and speaking tempo. *15th ICPhS*, 3129–3132.

Munro, M. J., & Derwing, T. M. (1995). Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners. *Language Learning*, *45*(1), 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Ordin, M., & Nespor, M. (2013). Transition Probabilities and Different Levels of Prominence in Segmentation. *Language Learning*, *63*(4), 800–834.

https://doi.org/10.1111/lang.12024

Ordin, M., Polyanskaya, L., Laka, I., & Nespor, M. (2017). Cross-linguistic differences in the use of durational cues for the segmentation of a novel language. *Memory & Cognition*, *45*, 863–876. https://doi.org/10.3758/s13421-017-0700-9

Osthoff, H., & Brugmann, K. (1878). Preface. In H. Osthoff & K. Brugmann (Eds.), *Morphologische Untersuchungen auf dem Gebiete der indogermanischen Sprachen, Volume 1*. Leipzig: Hirzel.

Pagel, M. (2017). Darwinian perspectives on the evolution of human languages. *Psychonomic Bulletin & Review*, *24*, 151–157. https://doi.org/10.3758/s13423-016-1072-z

Paul, H. (1880). *Prinzipien der Sprachgeschichte*. Halle: Max Niemeyer Verlag.

Petrone, C., Carbone, F., & Champagne-Lavau, M. (2016). Effects of emotional prosody on skin conductance responses in French. *Proceedings of the International Conference on Speech Prosody*, *2016-Janua*, 425–429. https://doi.org/10.21437/speechprosody.2016-87

Petrone, C., Truckenbrodt, H., Wellmann, C., Holzgrefe-lang, J., Wartenburger, I., & Höhle, B. (2017). Prosodic boundary cues in German : Evidence from the production and perception of bracketed lists. *Journal of Phonetics*, *61*, 71–92. https://doi.org/10.1016/j.wocn.2017.01.002

Pickering, M. J., & Branigan, H. P. (1999). Syntactic priming in language production. *Trends in Cognitive Sciences*, *3*(4), 136–141. https://doi.org/10.1016/S1364-6613(99)01293-0

Pisanski, K., Cartei, V., McGettigan, C., Raine, J., & Reby, D. (2016). Voice Modulation: A Window into the Origins of Human Vocal Control? *Trends in Cognitive Sciences*, *20*(4), 304–318. https://doi.org/10.1016/j.tics.2016.01.002

Piske, T., MacKay, I. R. A., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: a review. *Journal of Phonetics*, *29*, 191–215. https://doi.org/10.1006/jpho.2001.0134

Post, B., Marslen-Wilson, W. D., Randall, B., & Tyler, L. K. (2008). The processing of English regular inflections: Phonological cues to morphological structure. *Cognition*, *109*(1), 1–17. https://doi.org/10.1016/j.cognition.2008.06.011

Quam, C., & Swingley, D. (2012). Development in children's interpretation of pitch cues to emotions. *Child Development*, *83*(1), 236–250. https://doi.org/10.1111/j.1467-8624.2011.01700.x.Development

Rácz, P. (2013). *Salience in sociolinguistics*. Berlin: De Gruyter.

Ravignani, A., Barbieri, C., Flaherty, M., Jadoul, Y., Lattenkamp, E., Little, H., … Verhoef, T. (Eds.). (2020). The evolution of language - Proceedings of the 13th international conference. Nijmegen: The Evolution of Language Conferences.

Ravignani, A., Dalla Bella, S., Falk, S., Kello, C. T., Noriega, F., & Kotz, S. A. (2019). Rhythm in speech and animal vocalizations: a cross-species perspective. *Annals of the New York Academy of Sciences*, *1453*, 79–98. https://doi.org/10.1111/nyas.14166

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions : Relating regularization to inductive biases through iterated learning. *Cognition*, *111*(3), 317–328. https://doi.org/10.1016/j.cognition.2009.02.012

Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality and Social Psychology Review*, *8*(4), 364–382. https://doi.org/10.1207/s15327957pspr0804_3

Ritt, N. (2004). *Selfish sounds and linguistic evolution*. Cambridge: Cambridge University Press.

Roberts, G., & Fedzechkina, M. (2018). Social biases modulate the loss of redundant forms in the cultural evolution of language. *Cognition*, *171*, 194–201. https://doi.org/10.1016/j.cognition.2017.11.005

Robinson, P., & Ellis, N. C. (Eds.). (2008). *Handbook of Cognitive Linguistics and Second Language Acquisition*. New York: Routledge. https://doi.org/10.4324/9780203938560

Rosenbach, A. (2008). Variation, Selection, Development: Probing the Evolutionary Model of Language Change. In R. Eckardt, G. Jager, & T. Veenstra (Eds.), *Variation, Selection, Development: Probing the Evolutionary Model of Language Change* (pp. 23–74). Berlin: Mouton de Gruyter.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word Segmentation : The Role of Distributional Cues. *Journal of Memory and Language*, *35*, 606–621.

Scott, G. G., O'Donnell, P. J., & Sereno, S. C. (2012). Emotion words affect eye fixations during reading. *Journal of Experimental Psychology: Learning Memory and Cognition*, *38*(3), 783–792. https://doi.org/10.1037/a0027209

Seifart, F., Strunk, J., Danielsen, S., Hartmann, I., Pakendorf, B., Wichmann, S., … Bickel, B. (2021). The extent and degree of utterance-final word lengthening in spontaneous speech from 10 languages. *Linguistics Vanguard*, *7*(1), 1–14. https://doi.org/10.1515/lingvan-2019-0063

Seikel, J. A., Drumright, D. G., & Hudock, D. J. (Eds.). (2021). *Anatomy & physiology for speech, language, and hearing. Anatomy and physiology* (6th ed.). San Diego: Plural

Publishing.

Sereno, S. C., Scott, G. G., Yao, B., Thaden, E. J., & O'Donnell, P. J. (2015). Emotion word processing: does mood make a difference? *Frontiers in Psychology*, *6*(August), 1–13. https://doi.org/10.3389/fpsyg.2015.01191

Shattuck-Hufnagel, S., & Turk, A. E. (1996). A Prosody Tutorial for Investigators of Auditory Sentence Processing. *Journal of Psycholinguistic Research*, *25*(2), 193–247.

Singh, L., Reznick, J. S., & Xuehua, L. (2012). Infant word segmentation and childhood vocabulary development: a longitudinal analysis. *Developmental Science*, *15*(4), 482–495. https://doi.org/10.1111/j.1467-7687.2012.01141.x.Infant

Skov, M., & Nadal, M. (2020). The Nature of Beauty : behavior, cognition, and neurobiology. *Annals of the New York Academy of Sciences*, 1–12. https://doi.org/10.1111/nyas.14524

Slowiaczek, L. M. (1990). Effects of lexical stress in auditory word recognition. *Language and Speech*, *33*(1), 47–68. https://doi.org/10.1177/002383099003300104

Smith, K. (2006). Cultural Evolution of Language. *Encyclopedia of Language & Linguistics*, 315–322. https://doi.org/10.1016/B0-08-044854-2/04742-8

Smith, K. (2011). Learning bias, cultural evolution of language, and the biological evolution of the language faculty. *Human Biology*, *83*(2), 261–278. https://doi.org/10.3378/027.083.0207

Smith, K., Kirby, S., & Brighton, H. (2003). Iterated Learning: A Framework for the Emergence of Language. *Artificial Life*, *9*(4), 371–386. https://doi.org/10.1162/106454603322694825

Smith, K., Perfors, A., Fehér, O., Samara, A., Swoboda, K., & Wonnacott, E. (2017). Language learning, language use and the evolution of linguistic variation. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *372*(1711), 1–20.

Stefanowitsch, A. (2020). *Corpus linguistics: A Guide to the methodology*. Berlin: Language Sciences Press. Retrieved from http://langsci-press.org/catalog/book/000

Storkel, H. L. (2001). Learning new words: Phonotactic probability in language development. *Journal of Speech, Language, and Hearing Research*, *44*(6), 1321–1337. https://doi.org/10.1044/1092-4388(2001/103)

Styler, W. (2017). Using Praat for Linguistic Research. Retrieved from http://savethevowels.org/praat/

Tallerman, M., & Gibson, K. R. (Eds.). (2012). *The Oxford handbook of language evolution*. Oxford: Oxford University Press.

Tamariz, M., & Kirby, S. (2016). The cultural evolution of language. *Current Opinion in*

Psychology, *8*, 37–43. https://doi.org/10.1016/j.copsyc.2015.09.003

Taylor, A. M., & Reby, D. (2010). The contribution of source-filter theory to mammal vocal communication research. *Journal of Zoology*, *280*(3), 221–236. https://doi.org/10.1111/j.1469-7998.2009.00661.x

Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology*, *39*(4), 706–716. https://doi.org/10.1037/0012-1649.39.4.706

Tomasello, M., Kruger, A. C., & Ratner, H. H. (1993). Cultural learning. *Behavioral and Brain Sciences*, *16*(3), 495–552. https://doi.org/10.1017/s0140525x0003123x

Tyler, M. D., & Cutler, A. (2009). Cross-language differences in cue use for speech segmentation. *Journal of the Acoustical Society of America*, *126*(1), 367–376. https://doi.org/10.1121/1.3129127

Vaissière, J. (1983). Language-Independent Prosodic Features. In A. Cutler & D. R. Ladd (Eds.), *Springer Series in Language and Communication 14: Prosody: Models and Measurements* (pp. 53–66). Hamburg: Springer.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

Wellmann, C., Holzgrefe, J., Truckenbrodt, H., Wartenburger, I., & Höhle, B. (2012). How each prosodic boundary cue matters: Evidence from German infants. *Frontiers in Psychology*, *3*(DEC), 1–13. https://doi.org/10.3389/fpsyg.2012.00580

Yurovsky, D., Yu, C., & Smith, L. B. (2012). Statistical speech segmentation and word learning in parallel: Scaffolding from child-directed speech. *Frontiers in Psychology*, *3*(374), 1–9. https://doi.org/10.3389/fpsyg.2012.00374

Zarcone, A., van Schijndel, M., Vogels, J., & Demberg, V. (2016). Salience and attention in surprisal-based accounts of language processing. *Frontiers in Psychology*, *7*, 1–17. https://doi.org/10.3389/fpsyg.2016.00844

Zehentner, E. (2019). *Competition in language change - The rise of the English dative alternation*. Berlin: De Gruyter Mouton.

CHAPTER 1

Voice modulatory cues to structure across languages and species

This chapter is accepted for publication in *Philosophical Transactions of the Royal Society B*.

Citation:

Matzinger T & Fitch WT. Voice modulatory cues to structure across languages and species. *Philosophical Transactions of the Royal Society B* (accepted). https://doi.org/10.1098/rstb.2020-0393

Voice Modulatory Cues to Structure across Languages and Species

Theresa Matzinger
W Tecumseh Fitch

University of Vienna

Abstract

Voice modulatory cues such as variations in fundamental frequency, duration and pauses are key factors for structuring vocal signals in human speech and vocal communication in other tetrapods. Voice modulation physiology is highly similar in humans and other tetrapods due to shared ancestry and shared functional pressures for efficient communication. This has led to similarly structured vocalizations across humans and other tetrapods. Nonetheless, in their details, structural characteristics may vary across species and languages. Because data concerning voice modulation in non-human tetrapod vocal production and especially perception is relatively scarce compared to human vocal production and perception, this review focuses on voice modulatory cues used for speech segmentation across human languages, highlighting comparative data where available. Cues that are used similarly across many languages may help indicate which cues may result from physiological or basic cognitive constraints, and which cues may be employed more flexibly and are shaped by cultural evolution. This suggests promising candidates for future investigation of cues to structure in non-human tetrapod vocalizations.

Keywords

voice modulation, prosody, linguistic structure, cross-species comparisons, cross-linguistic comparisons, speech segmentation

1. Introduction

Although human speech is often thought to be categorically different from non-human animal vocal communication, many aspects of human acoustic communication are directly comparable with that of other land vertebrates. These include both the vocal apparatus itself and the main voice modulatory cues involved in vocal production.[1] In this review, we will argue that voice modulatory cues are similar in the vocal communication of humans and other tetrapods because of a) shared ancestry, resulting in a similar voice modulation physiology, and b) shared functional bases, i.e. similar pressures for efficient communication, resulting in similar cognitive processing due to domain-general mechanisms shared among species.

Voice modulatory cues that are shared and have similar functions in human and non-human tetrapod vocalizations as well as cross-linguistically can be hypothesized to result from anatomical, physiological and cognitive mechanisms that are evolutionarily conserved (1–3). These include vocal tract anatomy or respiratory constraints, along with domain-general learning constraints and/or cognitive production and perception constraints (e.g. attention and memory; 1,4,5). In contrast, cues that are neither paralleled in other tetrapods' vocalizations

[1] The terms "voice modulation" and "prosody" essentially describe the same concept, namely all kinds of vocal dynamic modifications of acoustic parameters during production in humans and non-human tetrapods (19,62,161). For the sake of consistency, we will use the term "voice modulation" throughout this review.

nor cross-linguistically varied may rely on less evolutionarily conserved mechanisms and therefore have larger potential to be shaped by cultural evolutionary processes. For example, the learnability and transmissibility of vocal features to future generations of signalers may not only be influenced by general mechanisms such as how easily the vocal features can be processed, but also by the social environment (6–11). Thus, factors such as group identity, community size or prestige may lead to different conventions of voice modulatory patterns in different communities (12,13). In this review, we attempt to begin disentangling which voice modulatory cues are the result of physiological constraints, of domain-general cognitive mechanisms, and of species- or language-specific conventions and learning pressures, aiming to contribute to the understanding of voice modulation in general evolutionary and cognitive terms.

Because this is a very large research program, our review will cover only some specific aspects of voice modulation. In the first section, we compare different voice modulatory cues across human speech and tetrapod communication, including pauses, fundamental frequency and syllable/unit duration. We discuss similarities and differences in the physiological mechanisms underlying these cues, and then discuss how the effort of producing and perceiving them may be linked to functional pressures in the environment. In the second part of the review, we take a comparative approach across languages, comparing if different voice modulatory cues used for speech segmentation are similar between or differ among various human languages. Especially regarding the many voice modulatory cues for which animal data remains scarce, comparisons between different human languages may provide valuable insights as to whether the physiological and cognitive mechanisms behind those cues are species-typical (and therefore may be evolutionarily conserved and domain-general), or more flexible language-specific. Finally, our review will identify research gaps and suggest avenues for further research that may help more clearly reveal the underlying physiological and cognitive mechanisms underlying the realization of different voice modulatory cues.

Overall, our comparison between voice modulatory cues in tetrapod vocalizations and across various human languages will show that biological evolution can constrain cultural evolution, and that many of the structures and cues widely used in human speech rely upon basic acoustic and cognitive mechanisms that humans share with other tetrapods.

2. Voice modulation physiology and constraints on vocal production

Humans and other tetrapods share many similarities in the physiological mechanisms used to produce vocal signals. Multiple similarities result from shared respiratory mechanisms, which in turn result from shared ancestry during biological evolution (14,15). Most tetrapods, including humans, produce vocal signals in a two-stage process: first, a **source** generates acoustic energy using an airflow from the lungs. This source is the larynx in most tetrapods and the syrinx in birds, and consists of vibrating tissue that creates sound by oscillating at a particular rate termed the fundamental frequency (f_o hereafter). This source signal is then **filtered** in the supralaryngeal vocal tract (upper respiratory tract) via multiple formant frequencies that act as a series of bandpass filters, attenuating or enhancing certain frequency ranges. The actual vocal output fuses these two components (source and filter), which are mostly independent, meaning that f_o can freely vary independent of formants and vice versa. This process, summarized as the source-filter-theory of vocal production, is shared by humans and most other tetrapods (16–19), with the exception of toothed whales (20) and certain whistle vocalization (e.g., in rodents; 21). This shared physiological basis of vocal production leads to many similarities in both the production and the acoustic output of humans and other tetrapods. Nonetheless, while constrained by physiological production mechanisms, voice modulatory cues can to a certain extent be flexible, and dynamic modifications of particular acoustic parameters can provide structure to the vocal output. Specific voice modulatory cues and the extent to which they can vary (Tab. 1) are reviewed below. In particular, we focus mainly on three cues that are well-investigated with regard to speech segmentation across human languages and will therefore be most relevant for the later sections of this review: pauses, pitch and durational cues.

Table 1. Voice modulatory cues in human and non-human tetrapod vocal signals, including the physiological factors which constrain them, and the specific ways in which they vary.

Shared voice modulatory cues in human and non-human tetrapod vocal signals	**constrained by**	**variation**
Pauses	lung capacities, respiration	number, duration, position
Fundamental frequency (pitch)	subglottal pressure, length of vibrating tissue	magnitude; location of modulation
Duration of syllables/units	lung capacities, respiration	magnitude; location of modulation
Intensity/amplitude (loudness)	effort with which air is pushed from the lungs	magnitude; location of modulation
Voice quality: formants, overtones and spectral envelope	physiology of the vocal tract, flexibility to move articulators	different sound qualities (timbre) and speech sounds (e.g. vowels)
Voice quality: glottal pulses	shape of the vibrating tissue, effort with which air is pushed from the lungs	manner of vibration and shape of the glottal pulses (e.g. breathy voice)

2.1 The physiology of pauses

Amongst the most distinctive voice modulatory cues are pauses in the vocal signal, which often result from the need to breathe via alternating between exhaling and inhaling. Typically, tetrapods vocalize during exhalation, and vocalizations pause during inhalation. However, some non-human tetrapods vocalize during both exhalation and inhalation, and thus do not need to pause during vocalization (e.g. donkey braying, chimpanzee pant hoots, or howler monkey howling, during which inhaling vocalizations are shorter than exhaling vocalizations, but similar in terms of structure and amplitude; 22). Humans are also capable of ingressive vocalizations such as gasps and chuckles, but these usually do not replace respiratory pauses and are less flexible in encoding meaning than egressive vocalizations (23–25). While pauses in tetrapods result from the same physiological mechanism, i.e. respiratory pausing, and are thus constrained by the individuals' lung capacities, they can also vary in their specific realizations. For example, pauses can differ in their duration, number, and their position in the vocal stream. Because of this flexibility, tetrapods, including humans, can use pauses to structure the vocal signal in many different ways (26). For example, birdsong is structured into units commonly termed "syllables" that are separated by short pauses during which rapid inhalation – "mini-breaths" – occur (27).

2.2 The physiology of duration

The duration of phonation at the source can induce durational and rate variations in the vocal output. These durational variations can extend over different domains of the vocal output, such as individual sounds, individual syllables/units or larger stretches of vocalizations (28,29). This can, for example, lead to different rhythmic patterns, to differences in vocalization tempo, or to distinctive vowel sounds in human speech, where phonemic distinctions between long and short durations are frequent. Duration of one syllable can also disambiguate neighboring phonemes, as exemplified in the American English words *ladder* (/æ/ longer) and *latter* (/æ / shorter), which only differ in their vowel length (30). Human speech sounds that differ in their vowel quality (determined by formants), such as the vowels in the English words *feet* and *fit*, may also have distinctive lengths. Again, physiologically, durational variations are limited by the individuals' breathing capacities, but below that capacious limit, duration can be varied more or less flexibly to give structure to the vocal output of humans and most non-human tetrapods alike.

2.3 The physiology of pitch

Vocal signals are further characterized by the vibration rate of the vibrating tissue, which determines the signals' f_o, often termed pitch in the speech literature (18). Typically, in tetrapods, f_o is influenced both by subglottal air pressure and by muscles that regulate the length and tension of the vibrating tissues, i.e. the vocal folds in non-avian tetrapods and the syringeal membranes in birds (31–33). By modulating these two factors, pitch can vary within and between vocal signals. To increase pitch, individuals can either increase the subglottal air pressure or the tension of the vibrating tissues. Both of these options require increased effort (see section 2), and can provide diversity and structure to vocal signals. For example, typically, on the level of syllables, an increase in pitch signals emphasis ('stress' in the speech literature), whereas pitch modulation on the phrase level can function as a boundary signal (34–37). Again, the effort required for pitch modulation, and physiology such as the dimensions of the vibrating tissues, limit the pitch range that can be realized. However, within that range, pitch can be employed flexibly to structure the vocal signal differently, as evidenced by different stress patterns observed in different languages (38).

Fundamentally, tetrapods share these voice modulatory cues because of their shared vocal production physiology, which in turn results from their shared ancestry. Nonetheless, the specific uses and manifestations of these cues can vary considerably across species and languages. For example, species, languages and individuals may differ in when and where they make pauses, when and where pitch rises and falls, or which segments they lengthen or shorten. One useful principle for categorizing and understanding this variation in vocal signals is based on the effort it takes to place emphasis in the vocal signal, using various voice modulatory cues. Thus, the following section will address emphasis and effort in the production of vocal signals, how they are influenced by functional pressures, and how this can lead to cultural evolution of prosodic patterns.

3. Emphasis and effort

It seems intuitively obvious that vocal signals can carry emphasis, and that this requires effort. In particular, it takes more effort to produce emphasized or stressed, i.e. louder, longer and higher pitched, syllables than non-emphasized or unstressed ones. However, despite a common assertion that producing certain voice modulatory cues is more "energetically efficient" than producing others (39–43), the exact metabolic costs needed to produce and process these cues have rarely been systematically compared. In fact, several studies have shown that vocalizing is not very costly in terms of oxygen, glucose or ATP needed (44–48). Thus, although it is clear that tensing muscles requires energy consumption, the costs involved in contracting the tiny muscles controlling source characteristics like f_0 may not be appreciable relative to an organism's overall energy budget. Respiratory muscles are larger and potentially more energy consuming, but they need to be constantly working to serve respiratory functions, independent of vocalization. The relative cost of increased vs. decreased pitch or duration during normal speech and frequently produced animal vocalizations will represent an even smaller proportion of net energy expenditure.[2] Finally, the cost of neuronal firing involved in producing or perceiving vocalizations is real, but also very difficult to quantify using current methods. Therefore, at present, we have little choice but to adopt an intuitive definition of "effort", which can manifest in dynamic effort, i.e. muscular effort for moving the articulators, and neural control effort, i.e. cognitive effort for planning, producing and processing voice modulatory cues. The term "stress" is used in phonology essentially as a catch-all term, connoting effort and emphasis, but not grounded in detailed syllable-by-syllable measures of expended effort.

How much effort senders will invest in emphasizing vocalizations is largely driven by an interplay of the functional pressure for successful versus efficient communication (40,49). These pressures may also influence which parts of the signal are emphasized. Emphasis can either extend over the whole signal (e.g. louder vocalizations in noisy environments) or be specific to certain elements of the signal (e.g. stressing certain phrases or syllables); the latter should be more energetically efficient, so we may expect organisms to vary cues across a vocal stream in many cases, as humans do with speech.

One well-studied example where signals are emphasized in their entirety is the so-called Lombard effect: both humans and other tetrapods, including non-human primates, birds and whales tend to vocalize louder and with a higher pitch, i.e. with an increased effort, when there is more background noise (45–49). When background noise in the environment is reduced, signalers return to vocalizations that need less effort and decrease their pitch and intensity. A recent example in birdsong occurred when traffic reductions during the Covid-19 pandemic resulted in lower-frequency bird vocalizations, showing that signalers can flexibly adapt their vocalizations to functional pressures in the environment (55).

Further examples of signals with emphasized elements include rhythmic vocalizations and stress or intonation patterns. This kind of emphasis needs both dynamic and cognitive effort on the side of the sender, but creates structure in the signal, which may reduce error, combat habituation, or facilitate meaning encoding and processing on the side of the listener (56). The complex interplay of pressures acting on the sender and receiver may lead to variation in vocal signals that is not fixed genetically but influenced by current properties of the environment (6,7) and shows that once individuals begin to produce vocal cues, there is an opportunity to modulate them. Furthermore, in species which learn their vocalizations (e.g. birdsong or human speech), small production or perception biases for or against certain voice modulatory structural patterns in a certain environment may be amplified over generations of speakers (6). This may lead to a process of cultural evolution, and can result in within-species variation in structural patterns of vocalizations as exemplified by different human languages or different dialects in other tetrapods' vocalizations (57,58).

[2] Note that respiratory muscles may induce higher energetic costs in very loud, high or long vocalizations such as such as during human singing and oratory, or mammalian roaring contests or infrasonic long-distance calls. Because subglottal pressure is an important factor determining both f_0 and sound intensity, very loud and high-pitched vocalizations may require more respiratory effort than normal breathing and vocalization. In addition, very long syllables may disrupt the natural respiratory rhythm.

Thus, overall, how exactly the different voice modulatory cues are utilized varies within physiological constraints, and results from a balancing act between communicating successfully, but with low effort. This in turn depends on functional pressures of listeners and environment, which can vary between different species and languages, and may include factors such as cultural evolution. How exactly different species and different linguistic communities deal with different functional pressures depends both on domain-specific factors such as auditory salience, domain-general cognitive constraints such as memory and attention, but also on more flexible constraints such as social factors. All of these factors will combine to constrain the range within which the different voice modulatory cues can be realized and determine the actual vocal output seen in a language or a species.

4. What we can learn from comparing voice modulatory cues across human languages

Different realizations of voice modulatory cues have been heavily investigated in human languages, but similar investigations in non-human tetrapod vocalizations are comparatively scarce and less systematic. Over the last decades, bioacoustics has made considerable advances in the investigation of non-human tetrapod vocal production, but research on the perception of voice modulatory cues in non-human tetrapods is still in its infancy (59,60). It is especially difficult to reach firm conclusions about the communicative meaning of voice modulatory structures found in non-human tetrapod vocal signals, given how few cues and species have been systematically investigated.

Therefore, the remaining sections of this review will mainly focus on the comparison of voice modulatory cues across human languages, and specifically the voice modulatory cues that help listeners to segment continuous speech into words. When voice modulatory cues are realized similarly across human languages, this suggests that fundamental physiological constraints or basic cognitive mechanisms may be responsible for these patterns (1–3), and that therefore, due to their shared ancestry, similar cues may also be prevalent in non-human tetrapod vocalizations. We suggest that such patterns may be provide starting points for investigating modulation in tetrapod vocal signals. In contrast, cues that differ across different linguistic communities may be largely influenced by different functional pressures in the environment and by cultural evolutionary processes, and therefore are more likely to also differ across tetrapod vocalizations.

Comparing voice modulation across human languages and non-human animal vocalizations, and using similarities and differences between them to draw conclusions about the evolutionary roots of vocal communication is not new (61–65). Similar approaches have already been proposed, for example, by Morton (63,64), who suggested that high and low pitch vocalizations signal similar emotions and attitudes across languages and species. Across species, a low pitch signals largeness, dominance and self-confidence whereas a high pitch signals smallness, submissiveness and prosociality. Ohala (66) suggests that this biological grounding helps to explain prosodic patterns that are consistent across human languages such as a final pitch decrease in declarative statements (i.e. utterances signaling dominance and self-assurance) and final pitch increase in questions (i.e. utterances signaling insecurity, submissiveness and need).

Past approaches typically either avoid detailing the specific acoustic cues (65), or treat these cues as fixed for a particular sound class (e.g. low-pitched growls and high-pitched whines). Our goal below is to call attention to how dynamics *within* a call can play a role in structuring acoustic signals, and to investigate the specific acoustic parameters varied. Furthermore, our approach extends previous proposals by highlighting the importance of listener-associated cognitive factors, such as perceptual salience, memory, attention and learnability of prosodic patterns, for biological and cultural evolution. Finally, our proposal captures a more diverse range of prosodic patterns than previous accounts. In contrast to Ohala (66), who explained prosodic patterns by primarily drawing on emotional communication, our account attempts to explain a more diverse set of linguistic structures and meanings.

5. Structure in human languages: the speech segmentation problem and cues to solving it

One crucial first step in the acquisition of linguistic structure is the segmentation of fluent speech into words, before the words' meaning is known. This so-called speech segmentation problem is most acute for infants learning their first language, but also concerns second language learners. For adults, the challenge is particularly evident when they try to identify distinct words while listening to an unfamiliar foreign language (67–69). Nevertheless, language learners eventually master the speech segmentation problem easily. This is because they implicitly use various cues in the speech stream to identify patterns and regularities, which in turn helps them to extract words. Such cues may also play a role in complex sequence learning in bird or whale song (e.g. 70), but this possibility remains little explored.

Speech segmentation is a challenge that speakers of all human languages have to face and that is therefore well suited for cross-linguistic comparisons. Over the last decades, cues used in human speech segmentation have been the subject of a large body of research in a variety of different languages such as English (71–75), German (76–79), Italian (77,78,80), French (73), Dutch (73), Spanish (78,81), Portuguese (82), Basque (78), Japanese (72), Cantonese, Mandarin and Russian (83). This makes it possible to compare the characteristics of speech segmentation cues across languages, answer questions about more general physiological and cognitive mechanisms that are necessary to create and process linguistic structure and identify functional pressures in the respective environments. Amongst the cues that have been identified to be very important for speech segmentation and creating linguistic structure are transitional probability cues ("statistical learning"), and the voice modulatory cues that are our focus (e.g. 67,72–74,84–91).

Transitional probability cues are based on listeners tracking the co-occurrence frequencies of syllables in vocal input (74,92; see 93 for a meta analysis). For example, when hearing the sound sequence *pretty#baby*, listeners can infer that *pretty* and *baby* are distinct words because the syllables *pre* and *ty* as well as *ba* and *by* also co-occur in other sequences such as *pretty#girl* or *lovely#baby*. In contrast, *ty* and *ba* co-occur less frequently and can therefore be assumed to span a word boundary (94). Speakers of a wide variety of languages have been demonstrated to use such transitional probability cues for language acquisition in similar ways (English: e.g. 71–75; German: 76–78; Italian: 77,78,80; French: 73; Dutch: 73; Spanish: 78,81; Portuguese: 82; Basque: 78; Japanese: 72). Notably, producing different speech sounds and syllable identities, is itself a form of voice modulation, and is a prerequisite for syllable creation and thus for tracking transitional probabilities. Specifically, individual vowels and consonants are created by moving the articulators, which leads to different formant frequency patterns (cf. Tab. 1; 95). While different languages have different speech sounds (38,96), the cross-linguistic ability to modulate the voice in a way that produces different speech sounds is crucial for the cross-linguistic use of transitional probabilities for speech segmentation.

Using transitional probabilities to infer characteristics of a signal appears to be a very general behavior since in basically any domain of action, including animal vocalizations, certain events are more likely to follow each other than others (97,98). In humans, the identification of transitional probability cues appears to be based on a domain-general cognitive mechanism, namely statistical learning (99–102). Furthermore, statistical learning is not a uniquely human cognitive mechanism, and also other species have been demonstrated to use it to deduce signal structure (103). These can even apply across species, for example many non-human animals form associations between heterospecific alarm calls and the presence of a predator (104,105). Also, vocal learning in non-human animals, most notably in birds, is suggested to be supported by statistical computations, although the precise mechanisms behind it are not yet fully understood (103). It thus seems likely that both humans and many non-human tetrapods rely on a combination of statistical learning and acoustic modulations when learning the structure of their species-specific sound sequences.

Statistical learning is a very general and prominent perceptual and cognitive skill. However, in human languages, voice modulatory cues in the speech stream, such as pauses, or variations in fundamental frequency, syllable duration or intensity (which create word stress, speech rhythm or intonation), can be processed more easily than statistical cues, and therefore have more significant effects on speech segmentation (67,75,79,80,90). However, since voice modulatory cues come in many different realizations and can have many different functions (106), their overall role in signaling linguistic structure, and the cognitive mechanisms needed for processing them, are less understood. While some voice modulatory cues are realized and processed similarly across languages (e.g. 72), others are subject to cross-linguistic variation (e.g. 73,78). This raises the question how much the realization and processing of voice modulatory cues is determined by domain-general cognitive or physiological constraints, and how much these cues may be shaped by cultural evolution.

6. Cues to speech perception: when voice modulatory cues count more than transitional probability cues

The efficiency of different voice modulatory cues for speech segmentation has traditionally been tested in artificial language learning experiments (74). In these experiments, participants are exposed to several minutes of a continuous stream of nonsense speech, consisting of randomly concatenated invented pseudo-words. Listeners can infer from the transition probabilities between syllables which syllable combinations are “words” of the artificial language and can segment these items from the stream. To test the influence of voice modulatory cues on listeners’ segmentation performance, voice modulatory cues are added at different positions to the speech stream and it is measured how this changes listeners’ perception of words in the stream.

In such artificial language learning experiments, voice modulatory cues added to continuous speech on the word (e.g. 72,73,79) and phrase level (e.g. 107–110) typically enhance speech segmentation compared to transitional probability cues only. Crucially, these cues facilitate speech segmentation most effectively when they converge with the transitional probability cues in the speech stream, i.e. when the voice modulatory cues sound as “natural” to the listeners as they do in natural speech. In contrast, when voice modulatory cues are designed to conflict with the transitional probability cues in experimental settings and sound “unnatural” to the listeners, voice modulatory cues disrupt speech segmentation or even override the transitional probability cues (67,75,79,80,90). Whether voice modulatory cues at certain positions in the speech stream sound natural or unnatural with respect to the transitional probability cues depends both on language-universal cognitive predispositions such as attention, perception or preferences in pattern recognition, and on language-specific word stress patterns typical of the listeners’ native languages (72,73,80).

Crucially, many artificial language learning studies tested the influence of language-specific word stress on speech segmentation by using a combination of different voice modulatory cues (73,77,80). For example, stress cues dominated transitional probability cues when they were implemented as a combination of longer duration, higher pitch and higher intensity of stressed syllables (67,75,90). While using a combination of different voice modulatory cues closely simulates natural languages (69,90,91), it does not tell anything about the effects of the individual voice modulatory cues in isolation. However, since different voice modulatory cues have different physiological origins and may be cognitively processed and culturally transmitted differently, investigating them separately can reveal more about the functional pressures acting on linguistic structure (80,87).

Several studies have already addressed the role of voice modulatory cues in isolation. These studies suggest that pauses and lengthening serve as language-universal signals for word-finality (e.g. 72,73,77,78,84,87,111; but also: 80,112). In contrast, pitch increase is suggested to be the main perceptual correlate of word stress and is therefore processed differently by speakers of different languages (67,73,77,113). Speech segmentation studies investigating other prosodic cues such as intensity or voice quality are comparatively rare (87,114), which is why our review below focuses on pauses, durational and pitch modifications.

7. Pauses

Pause cues typically result from the physiological necessity to breathe, but pauses could in principle be expressed at different positions in a vocal signal, or differ in number and duration. Still, in practice, pauses are realized in strikingly similar ways across human languages. Language-universally, pauses are realized at the end of sentences or phrases but hardly ever occur within phrases or within words (26,115). This is further supported by second language learning studies finding that second language learners have hardly any problems acquiring pause characteristics typical of their second language (116,117). Thus, while in principle, pauses could occur anywhere within the breathing range, it is most probable that domain-general cognitive processing mechanisms constrain them to occur at specific positions in the vocal output – namely at those positions where they structure the vocal output most efficiently and with the least processing effort.

This, and their perceptual salience may explain why pauses are very effective for speech segmentation and outrank other cues in speech segmentation experiments (79).

In animal vocal signals it is challenging to determine whether pauses occur between or within phrases because units and phrases in animal vocalizations are less clearly defined (118). Still, because of their shared ancestry with humans, it can be expected that in non-human tetrapods' vocalizations, pauses manifest similarly, i.e. at the end of phrases or units. This is why pauses are often used by researchers to determine units in non-human tetrapod vocalizations (119).

8. Final lengthening as a cross-linguistic segmentation cue

One reason why final lengthening may serve as a language-independent speech segmentation cue is that language-universally, sentence-final or phrase-final elements are lengthened in everyday speech production (26,73,120–123). The evolutionary origins of final lengthening are that at sentence or phrase boundaries, speakers need to switch from exhaling to inhaling, leading to a pause, and that it takes less effort to slow articulators down before a pause than to stop them abruptly (124–128). Similar patterns can also be observed in movements in other domains than vocalization. For example, also runners decelerate their movements before stopping (129). This mechanistic factor seems like a good candidate for a factor that could play a role across languages, and in other species' vocal communication systems: a potential universal in vocal communication.

Because kinematic articulatory constraints result in lengthened syllables before sentence or phrase boundaries, listeners may have learned to associate lengthening with boundaries and to exploit it as a cue for speech segmentation (130). In turn, speakers may have started to intentionally use lengthening to indicate boundaries in the speech stream, also at positions where they did not pause (131). Via cultural transmission, this may have resulted in final lengthening becoming a conventionalized but still language-universal boundary signal (132). Because final lengthening is used as a convention for indicating boundaries cross-linguistically, it can be assumed that besides the articulatory constraints that speakers of all languages face equally, its transmission and processing is based on domain-general cognitive constraints.

This notion is supported by the putatively language-independent Iambic/Trochaic Law (= ITL; 133–137), which states that cross-linguistically, listeners group sounds with longer duration as sequence-final (iambic grouping). Although the ITL focuses on disyllabic words, it can also be generalized to trisyllabic words, suggesting that domain-general cognitive mechanisms may be responsible for this flexibility (72,79). Still, recently, there has also been evidence that the perceptual groupings of sequences of syllables with variable duration may be shaped more by cultural variation than previously assumed (80,138–140). Interestingly, the ITL not only applies to linguistic stimuli, but also to tone sequences (114,136) or visual patterns (141). This further supports the idea that final lengthening as a signal to linguistic structure and thus to low-effort communication results from general cognitive processing mechanisms that also apply to non-linguistic stimuli.

Since deceleration before pauses occurs across various human movements (129) and final lengthening is perceived as a boundary signal across different sensory domains, the mechanisms behind it seem likely to be evolutionarily old. Because of their shared ancestry with humans, a similar vocal tract physiology and similar energetic constraints, final lengthening and its perception as a boundary signal are promising targets for investigation in non-human tetrapods, and there is already some evidence for final lengthening in birdsong (142,143). Such a cue could play an important role, for example, in structuring turn-taking exchanges between individuals (144,145). However, to our knowledge, there is no current evidence that non-human tetrapods use final lengthening as a boundary cue at a perceptual level, and when listening to human speech, rats do not appear to group syllables varying in duration according to the ITL (137). Research with other tetrapods is badly needed to further examine this potential universal.

9. Pitch cues as language-specific segmentation cues

In multiple speech segmentation experiments, similar pitch modifications led to different segmentation patterns in speakers of different native languages (73,77). For example, word-initial pitch increase facilitated speech segmentation for native speakers of English, whereas word-final pitch increase facilitated speech segmentation for native speakers of French. These patterns are consistent with the typical stress placements of these languages (73,146).

One explanation why duration and pitch are used differently for speech segmentation is that, potentially, pitch is used as a more reliable cue for the perception of word stress than duration.

In speech *production*, stressed syllables are characterized by a co-occurrence of higher pitch and longer duration, and interestingly, cross-linguistically, duration seems to be a more consistent marker of word stress than pitch (80,147; but also: 73 for French and English). Still, while being an important acoustic correlate of word stress, lengthening at the same time occurs at boundaries (as discussed in the previous section) and most likely, this durational increase is larger and more consistently applied than that at stressed syllables (124). As a result, during *perception*, to avoid ambiguities, listeners may rely on lengthening for perceiving boundaries, but rather focus on pitch for perceiving word stress (73,79).

In general, listeners may need to be more flexible in the perception and cognitive processing of pitch variations compared to durational variations. In natural speech, pitch as a signal for word stress varies more than duration as a signal for sentence or phrase finality, for example because of loan words with non-typical stress patterns (148–150). In addition, intonation patterns are variable and depend for example on speaker emotions, attitudes, grammatical structure and focus (151). Also, while sentence-final pitch decrease in declarative sentences is common across languages (109,122,152), listeners may equally encounter sentence-final pitch increase in yes-no questions. Therefore, overall, pitch may be a less consistent (39,153–155) and less informative cue during speech segmentation than lengthening. This may explain why neither word-final pitch decrease (79) nor increase facilitated speech segmentation (73,77,156) in artificial language learning experiments, unless for speakers of languages with word-final stress (73,146).

According to the ITL (135,137,157–159), listeners perceive sounds with a higher pitch in sequence-initial positions (trochaic grouping). Interestingly, rats similarly group sequences that vary in pitch as trochees (137). However, apparently, this perceptual grouping does not play a big role for speech segmentation, since cross-linguistically, a word-initial higher pitch has facilitated speech segmentation in artificial language learning experiments only inconsistently (73,79,80,156). It can therefore be inferred that the ITL for pitch does not systematically generalize from disyllabic to trisyllabic words, but pitch is instead processed more flexibly.

The apparently rather flexible processing of pitch may result in weak production, perception or learning biases amplifying pitch cues in different directions during the cultural transmission of languages. This may in turn lead to different stress patterns in different languages, making pitch a less reliable signal for speech segmenation than duration. While still originating from basic cognitive processing mechanisms, the cognitive and physiological structures responsible for pitch processing are therefore suggested to be less conserved than those responsible for duration processing. This may have constrained the cultural evolution of pitch cues to linguistic structure less than that of durational cues. Thus, functional pressures for structured signals may hold equally across languages, but how exactly this linguistic structure is archieved, can vary cross linguistically.

While lexical stress patterns vary across languages and it can be assumed that similar variation should be expected in other tetrapod vocalizations, utterance-final pitch decrease in declarative statements is common across many languages (37,109,122,152). One reason for this declination may be that the articulators, in this case the vibrating tissues, are slowed down before being brought to a halt, and this lower vibration rate of the tissues leads to a lower pitch (160). A functional reason may be that pitch declination facilitates turn taking and thus decreases communicative effort[3]. These physiological and functional constraints are shared across species, which is why pitch declination may be an interesting target for investigation in non-human tetrapod vocal signals. Indeed, there are some indications for final pitch declination and turn taking in vervet monkeys and rhesus macaques (36). Investigating other species for final pitch declination could further corroborate the hypothesis that a shared ancestry drives similarities in pitch realization and processing in humans and non-human tetrapods.

10. Conclusions and outlook

Summarizing, our review of human speech modulation shows that *fo*, duration and pauses are typically used in systematic ways across languages to help structure the speech signal, but that there is nonetheless considerable variation across languages in the details. Voice modulation can, in many cases, provide cues to structure that are more salient and effective to listeners and learners than statistical measures over the vocal units (e.g. sequential transition probabilities), and can work together with such statistical information or in some cases override it. Thus, although such statistical cues are important (and can be readily computed in animal signals like bird or whale song), they obscure the importance of voice modulation as a key factor in structuring animal communication signals.

How language- or species-specific and cross-linguistic and cross-species cues interact certainly warrants further research. In those cases where comparative information is available, it suggests that the cues used to indicate structure in the speech signal are both present in vocalizations of other species (unsurprising given their fundamentally similar production mechanisms) and also can be used in similar ways (e.g. phrase final lengthening in speech and birdsong). Nonetheless, there is currently far too little comparative data to allow any clear conclusions about the degree to which human-typical cues to structure are also used by other species. More research in this area – what we might term "animal phonology" – is needed to evaluate whether there are broad phylogenetic generalizations to be made, as we have hypothesized here. A rich comparative analysis of these issues could be expected to shed light not just on the evolution of communication across vertebrates, but also about the phylogenetic origins of universals in human speech production and perception.

Acknowledgements

We would like to thank Andrey Anikin and an anonymous reviewer for helpful comments on an earlier version of this manuscript.

Funding

This work was supported by the Austrian Science Fund (FWF) DK Grant Cognition & Communication [grant number: #W1262-B29] to WTF.

Author contributions

TM: Conceptualization, Writing – original draft; WTF: Conceptualization, Writing – review and editing, Supervision

[3] However, potential analogies between turn taking in human and non-human animal vocalizations have to be interpreted with caution. Since it is difficult to assess the underlying meaning or the intentions behind non-human animal vocal signals, alternation of signals may not necessarily be the result of active turn-taking (145). In such cases, the communicative benefit gained from alternating vocalizations may differ among species.

References

1. Fitch WT. The Evolution of Language. Cambridge University Press; 2010.
2. ten Cate C. Assessing the uniqueness of language: Animal grammatical abilities take center stage. Psychon Bull Rev. 2017;24(1):91–6.
3. Christiansen MH, Chater N. The language faculty that wasn't: A usage-based account of natural language recursion. Front Psychol. 2015;6(AUG):1–18.
4. Evans N, Levinson SC. The myth of language universals : Language diversity and its importance for cognitive science. Behav Brain Sci. 2009;(32):429–92.
5. Fitch WT, Boer B de, Mathur N, Ghazanfar AA. Monkey vocal tracts are speech-ready. Sci Adv [Internet]. 2016 Dec 12;2(12):e1600723. Available from: http://advances.sciencemag.org/content/2/12/e1600723
6. Smith K. Learning bias, cultural evolution of language, and the biological evolution of the language faculty. Hum Biol. 2011;83(2):261–78.
7. Smith K, Kirby S. Cultural evolution: implications for understanding the human language faculty and its evolution. Philos Trans R Soc Lond B Biol Sci. 2008;363(1509):3591–603.
8. Smith K, Kalish ML, Griffiths TL, Lewandowsky S. Introduction. Cultural transmission and the evolution of human behaviour. Philos Trans R Soc B Biol Sci. 2008;363(1509):3469–76.
9. Watson SK, Townsend SW, Schel AM, Wilke C, Wallace EK, Cheng L, et al. Vocal learning in the functionally referential food grunts of chimpanzees. Curr Biol. 2015;25(4):495–9.
10. Whiten A. Cultural Evolution in Animals. Annu Rev Ecol Evol Syst. 2019;50:27–48.
11. Williams H, Levin II, Norris DR, Newman AEM, Wheelwright NT. Three decades of cultural evolution in Savannah sparrow songs. Anim Behav [Internet]. 2013;85(1):213–23. Available from: http://dx.doi.org/10.1016/j.anbehav.2012.10.028
12. Smith K, Perfors A, Fehér O, Samara A, Swoboda K, Wonnacott E. Language learning, language use and the evolution of linguistic variation. Philos Trans R Soc London B Biol Sci. 2017;372(1711):1–20.
13. Raviv L, de Heer Kloots M, Meyer A. What makes a language easy to learn? A preregistered study on how systematic structure and community size affect language learnability. Cognition [Internet]. 2021;210(February):104620. Available from: https://doi.org/10.1016/j.cognition.2021.104620
14. Hoffman M, Taylor BE, Harris MB. Evolution of lung breathing from a lungless primitive vertebrate. Respir Physiol Neurobiol. 2016;224:11–6.
15. Perry SF, Sander M. Reconstructing the evolution of the respiratory apparatus in tetrapods. Respir Physiol Neurobiol. 2004;144(2-3 SPEC. ISS.):125–39.
16. Titze I. Principles of voice production. Englewood Cliffs: Prentice Hall; 1994.
17. Taylor AM, Reby D. The contribution of source-filter theory to mammal vocal communication research. J Zool. 2010;280(3):221–36.
18. Fitch WT. The evolution of speech: a comparative review. Trends Cogn Sci [Internet]. 2000;4(7):258–67. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10859570
19. Pisanski K, Cartei V, McGettigan C, Raine J, Reby D. Voice Modulation: A Window into the Origins of Human Vocal Control? Trends Cogn Sci. 2016;20(4):304–18.
20. Reidenberg JS, Laitman JT. Anatomy of Underwater Sound Production With a Focus on Ultrasonic Vocalization in Toothed Whales Including Dolphins and Porpoises. In: Handbook of Behavioral Neuroscience. Elsevier B.V.; 2018. p. 509–19.
21. Riede T, Borgard HL, Pasch B. Laryngeal airway reconstruction indicates that rodent ultrasonic vocalizations are produced by an edge-tone mechanism. R Soc Open Sci. 2017;4(11).

22. de Cunha RGT, de Oliveira DAG, Holzmann I, Kitchen DM. Production of loud and quiet calls in Howler Monkeys. Kowalewski MM, Garber PA, Cortés-Ortiz L, Urbani B, Youlatos D, editors. Howler Monkeys: Adaptive Radiation, Systematics, and Morphology. New York: Springer; 2015. 337–368 p.
23. Eklund R. Pulmonic ingressive phonation: Diachronic and synchronic characteristics, distribution and function in animal and human sound production and in human speech. J Int Phon Assoc. 2008;38(3):235–324.
24. Eklund R. Pulmonic ingressive speech: a neglected universal? Proc Fonetik. 2007;50(August):21–4.
25. Eklund R. Pulmonic ingressive speech. In: Damico JS, Ball MJ, editors. The SAGE Encyclopedia of Human Communication Sciences and Disorders. Thousand Oaks, CA: Sage Publications; 2019. p. 1529–32.
26. Fletcher J. The Prosody of Speech : Timing and Rhythm. In: Hardcastle WJ, Laver J, Gibbon FE, editors. The handbook of phonetic sciences. 2nd ed. Hoboken: Wiley-Blackwell; 2010. p. 523–602.
27. Hartley RS, Suthers RA. Airflow and pressure during canary song: direct evidence for mini-breaths. J Comp Physiol A. 1989;165:15–26.
28. Ravignani A, Dalla Bella S, Falk S, Kello CT, Noriega F, Kotz SA. Rhythm in speech and animal vocalizations: a cross-species perspective. Ann N Y Acad Sci. 2019;1453:79–98.
29. Jacewicz E, Fox RA, Wei L. Between-speaker and within-speaker variation in speech tempo of American English. J Acoust Soc Am [Internet]. 2010;128(2):839–50. Available from: http://asa.scitation.org/doi/10.1121/1.3459842
30. House AS. On Vowel Duration in English. J Acoust Soc Am. 1961;33(9):1174–8.
31. Riede T. Subglottal pressure, tracheal airflow, and intrinsic laryngeal muscle activity during rat ultrasound vocalization. J Neurophysiol. 2011;106(5):2580–92.
32. Riede T, Tokuda IT, Farmer CG. Subglottal pressure and fundamental frequency control in contact calls of juvenile Alligator mississippiensis. J Exp Biol. 2011;214(18):3082–95.
33. Riede T, Goller F. Functional morphology of the sound-generating labia in the syrinx of two songbird species. J Anat. 2010;216(1):23–36.
34. Lehiste I. Suprasegmentals. Cambridge: MIT Press; 1970.
35. Adams C, Munro RR. In search of the acoustic correlates of stress: fundamental frequency. Phonetica [Internet]. 1978;125–56. Available from: http://marefateadyan.nashriyat.ir/node/150
36. Hauser MD, Fowler CA. Fundamental frequency declination is not unique to human speech: Evidence from nonhuman primates. J Acoust Soc Am. 1992;91(1):363–9.
37. Pierrehumbert J. The perception of fundamental frequency declination. J Acoust Soc Am. 1979;66(2):363–9.
38. Dryer MS, Haspelmath M. The World Atlas of Language Structures Online [Internet]. Dryer MS, Haspelmath M, editors. Munich: Max Planck Digital Library; 2013. Available from: https://wals.info
39. Bybee J. Frequency of use and the organization of language. Oxford: Oxford University Press; 2007.
40. Fedzechkina M, Jaeger TF, Newport EL. Language learners restructure their input to facilitate efficient communication. Proc Natl Acad Sci U S A. 2012;109(44):17897–902.
41. Gibson E, Futrell R, Piandadosi ST, Dautriche I, Mahowald K, Bergen L, et al. How Efficiency Shapes Human Language. Trends Cogn Sci. 2019;23(5):389–407.
42. Blevins J. Evolutionary Phonology - The Emergence of Sound Patterns. Vol. 112. Cambridge: Cambridge University Press; 2004.

43. Zipf GK. Human behavior and the principle of least effort. Cambridge, MA: Addison-Wesley Press; 1949.
44. Moon S-J, Lindblom B. Two experiments on oxygen consumption during speech production: vocal effort and speaking tempo. 15th ICPhS. 2003;3129–32.
45. Horn AG, Leonard ML, Weary DM. Oxygen consumption during crowing by roosters: talk is cheap. Anim Behav. 1995;50(5):1171–5.
46. Oberweger K, Goller F. The metabolic cost of birdsong production. J Exp Biol. 2001;204(19):3379–88.
47. Speakman JR, Racey PA. No cost of echolocation for bats in flight. Nature. 1991;350(6317):421–3.
48. Foskolos I, Aguilar de Soto N, Madsen PT, Johnson M. Deep-diving pilot whales make cheap, but powerful, echolocation clicks with 50 µL of air. Sci Rep [Internet]. 2019;9(1):1–9. Available from: http://dx.doi.org/10.1038/s41598-019-51619-6
49. Grice HP. Logic and conversation. In: Cole P, Morgan J, editors. Syntax and semantics. New York: Academic Press; 1975. p. 41–58.
50. Brumm H, Naguib M. Chapter 1 Environmental Acoustics and the Evolution of Bird Song. Adv Study Behav. 2009;40(December 2009):1–33.
51. Brumm H, Zollinger A. The evolution of the Lombard effect: 100 years of psychoacoustic research. Behaviour. 2011;148(11–13):1173–98.
52. Egnor SER, Hauser MD. Noise-induced vocal modulation in cotton-top tamarins (Saguinus oedipus). Am J Primatol [Internet]. 2006;68:1183–90. Available from: https://www.rgs.org/NR/rdonlyres/2C503639-2BCE-4580-AA66-8725DA4D412A/0/BatManualUpdated.pdf
53. Nemeth E, Pieretti N, Zollinger SA, Geberzahn N, Partecke J, Mirand AC, et al. Bird song and anthropogenic noise: Vocal constraints may explain why birds sing higher-frequency songs in cities. Proc R Soc B Biol Sci. 2013;280(1754):1–7.
54. Manabe K, Sadr EI, Dooling RJ. Control of vocal intensity in budgerigars (*Melopsittacus undulatus*): Differential reinforcement of vocal intensity and the Lombard effect. J Acoust Soc Am [Internet]. 1998;103(2):1190–8. Available from: http://asa.scitation.org/doi/10.1121/1.421227
55. Derryberry EP, Phillips JN, Derryberry GE, Blum MJ, Luther D. Singing in a silent spring: Birds respond to a half-century soundscape reversion during the COVID-19 shutdown. Science (80-). 2020;370(6516):575–9.
56. Jaeger TF, Tily H. On language “utility”: Processing complexity and communicative efficiency. Wiley Interdiscip Rev Cogn Sci. 2011;2(3):323–35.
57. Garland EC, Goldizen AW, Rekdahl ML, Constantine R, Garrigue C, Hauser ND, et al. Dynamic horizontal cultural transmission of humpback whale song at the ocean basin scale. Curr Biol [Internet]. 2011;21(8):687–91. Available from: http://dx.doi.org/10.1016/j.cub.2011.03.019
58. Laland KN, Galef B, editors. The question of animal culture. Cambridge, MA: Harvard University Press; 2009.
59. Bradbury JW, Vehrencamp SL. Principles of animal communication. 2nd ed. Sunderland, MA: Sinauer Associates; 2011.
60. Erbe C, Dent ML. Animal Bioacoustics. Acoust Today. 2017;13(2):65–7.
61. Filippi P, Hoeschele M, Spierings M, Bowling DL. Temporal modulation in speech, music, and animal vocal communication: evidence of conserved function. Ann N Y Acad Sci. 2019;1453(1):99–113.
62. Filippi P. Emotional and interactional prosody across animal communication systems: A comparative approach to the emergence of language. Front Psychol. 2016;7(SEP):1–19.
63. Morton ES. On the Occurrence and Significance of Motivation-Structural Rules in

Some Bird and Mammal Sounds. Am Nat. 1977;111(981):855–69.
64. Morton ES. Grading, discretenesss, redundancy, and motivational-structural rules. In: Kroodsma D, Miller EH, editors. Acoustic communication in birds. New York: Academic Press; 1982. p. 183–212.
65. Filippi P, Congdon J V., Hoang J, Bowling DL, Reber SA, Pašukonis A, et al. Humans recognize emotional arousal in vocalizations across all classes of terrestrial vertebrates: Evidence for acoustic universals. Proc R Soc B Biol Sci. 2017;284(20170990):1–9.
66. Ohala JJ. Cross-language use of pitch: an ethological view. Phonetica. 1983;40:1–18.
67. Johnson EK, Jusczyk PW. Word Segmentation by 8-Month-Olds : When Speech Cues Count More Than Statistics. J Mem Lang. 2001;44:548–67.
68. Endress AD, Hauser MD. Word segmentation with universal prosodic cues. Cogn Psychol [Internet]. 2010;61(2):177–99. Available from: http://dx.doi.org/10.1016/j.cogpsych.2010.05.001
69. Erickson LC, Thiessen ED. Statistical learning of language: Theory, validity, and predictions of a statistical learning account of language acquisition. Dev Rev [Internet]. 2015;37:66–108. Available from: http://dx.doi.org/10.1016/j.dr.2015.05.002
70. Fehér O, Ljubičić I, Suzuki K, Okanoya K, Tchernichovski O. Statistical learning in songbirds: From selftutoring to song culture. Philos Trans R Soc B Biol Sci. 2017;372(1711).
71. Mattys SL, Jusczyk PW, Luce PA, Morgan JL. Phonotactic and Prosodic Effects on Word Segmentation in Infants. Cogn Psychol. 1999;38(4):465–94.
72. Frost RLA, Monaghan P, Tatsumi T. Domain-general mechanisms for speech segmentation: The role of duration information in language learning. J Exp Psychol Hum Percept Perform. 2017;43(3):466–76.
73. Tyler MD, Cutler A. Cross-language differences in cue use for speech segmentation. J Acoust Soc Am. 2009;126(1):367–76.
74. Saffran JR, Aslin RN, Newport EL. Statistical learning by 8-month-old infants. Science (80-). 1996;274(5294):1926–8.
75. Thiessen ED, Saffran JR. When cues collide: use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. Dev Psychol [Internet]. 2003;39(4):706–16. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12859124
76. Marimon Tarter M. Word segmentation in German-learning infants and German-speaking adults: prosodic and statistical cues. University of Potsdam; 2019.
77. Ordin M, Nespor M. Native Language Influence in the Segmentation of a Novel Language. Lang Learn Dev. 2016;12(4):461–81.
78. Ordin M, Polyanskaya L, Laka I, Nespor M. Cross-linguistic differences in the use of durational cues for the segmentation of a novel language. Mem Cognit. 2017;45:863–76.
79. Matzinger T, Ritt N, Fitch WT. The Influence of Different Prosodic Cues on Word Segmentation. Front Psychol. 2021;12.
80. Ordin M, Nespor M. Transition Probabilities and Different Levels of Prominence in Segmentation. Lang Learn. 2013;63(4):800–34.
81. Cunillera T, Càmara E, Laine M, Rodríguez-Fornells A. Speech segmentation is facilitated by visual cues. Q J Exp Psychol. 2009;63:1–15.
82. Fernandes T, Ventura P, Kolinsky R. The Relative Weight of Statistical and Prosodic Cues in Speech Segmentation: A Matter of Language-(In)dependency and of Signal Quality. J Port Linguist. 2011;10(1):87.
83. Gómez DM, Mok P, Ordin M, Mehler J, Nespor M. Statistical Speech Segmentation in Tone Languages: The Role of Lexical Tones. Lang Speech. 2018;61(1):84–96.
84. Saffran JR, Newport EL, Aslin RN. Word Segmentation : The Role of Distributional Cues. J Mem Lang. 1996;35:606–21.

85. Saffran JR, Johnson EK, Aslin RN, Newport EL. Statistical learning of tone sequences by human infants and adults. Cognition [Internet]. 1999 [cited 2016 Dec 16];70(1):27–52. Available from: http://www.sciencedirect.com/science/article/pii/S0010027798000754
86. Aslin RN, Saffran JR, Newport EL. Computation of conditional probability statistics by human infants. Psychol Sci. 1998;9(4):321–4.
87. Hay JSF, Saffran JR. Rhythmic grouping biases constrain infant statistical learning. Infancy. 2012;17(6):610–41.
88. Johnson EK. Infants use prosodically conditioned acoustic-phonetic cues to extract words from speech. J Acoust Soc Am [Internet]. 2008;123(6):EL144–8. Available from: http://asa.scitation.org/doi/10.1121/1.2908407
89. Johnson EK. Bootstrapping language : Are infant statisticians up to the job ? In: Rebuschat P, Williams J, editors. Statistical learning and language acquisition. Berlin: Mouton de Gruyter; 2012. p. 55–90.
90. Johnson EK, Seidl AH. At 11 months, prosody still outranks statistics. Dev Sci. 2009;12(1):131–41.
91. Johnson EK, Tyler MD. Testing the Limits of Statistical Learning for Word Segmentation. Dev Sci. 2010;13(2):339–45.
92. Romberg AR, Saffran JR. Statistical learning and language acquisition. Wiley Interdiscip Rev Cogn Sci. 2010;1(6):906–14.
93. Black A, Bergmann C. Quantifying infants' statistical word segmentation: A meta-analysis. Proc 39th Annu Conf Cogn Sci Soc [Internet]. 2017;(3):124–9. Available from: https://pdfs.semanticscholar.org/0807/41051b6e2b74d2a1fc2e568c3dd11224984b.pdf
94. Saffran JR. Statistical language learning: Mechanisms and constraints. Curr Dir Psychol Sci. 2003;12(4):110–4.
95. Redford MA, editor. The Handbook of Speech Production. Chichester: Wiley Blackwell; 2015.
96. Moran S, McCloy D. PHOIBLE 2.0. 2019.
97. Gagniuc PA. Markov Chains: From Theory to Implementation and Experimentation. Hoboken: John Wiley & Sons; 2017.
98. Shannon CE. A Mathematical Theory of Communication. Bell Syst Tech J. 1948;27(4):623–56.
99. Bogaerts L, Frost R, Christiansen MH. Integrating statistical learning into cognitive science. J Mem Lang [Internet]. 2020;115(August):104167. Available from: https://doi.org/10.1016/j.jml.2020.104167
100. Newport EL. Statistical language learning: computational, maturational, and linguistic constraints. Lang Cogn. 2016;8(3):447–61.
101. Thiessen ED, Erickson LC. Beyond Word Segmentation: A Two- Process Account of Statistical Learning. Curr Dir Psychol Sci. 2013;22(3):239–43.
102. Palmer SD, Mattys SL. Speech segmentation by statistical learning is supported by domain-general processes within working memory. Q J Exp Psychol. 2016;69(12):2390–401.
103. Santolin C, Saffran JR. Constraints on Statistical Learning Across Species. Trends Cogn Sci [Internet]. 2018;22(1):52–63. Available from: http://dx.doi.org/10.1016/j.tics.2017.10.003
104. Hauser MD. How infant vervet monkeys learn to recognize starling alarm calls: the role of experience. Behaviour. 1988;105:187–201.
105. Rainey HJ, Zuberbühler K, Slater PJB. Hornbills can distinguish between primate alarm calls. Proc R Soc B Biol Sci. 2004;271:755–9.
106. Cole J. Prosody in context: a review. Lang Cogn Neurosci. 2015;30(1–2):1–31.

107. Christophe A, Peperkamp S, Pallier C, Block E, Mehler J. Phonological phrase boundaries constrain lexical access I. Adult data. J Mem Lang. 2004;51(4):523–47.
108. Gout A, Christophe A, Morgan JL. Phonological phrase boundaries constrain lexical access II. Infant data. J Mem Lang. 2004;51(4):548–67.
109. Langus A, Marchetto E, Bion RAH, Nespor M. Can prosody be used to discover hierarchical structure in continuous speech? J Mem Lang [Internet]. 2012;66(1):285–306. Available from: http://dx.doi.org/10.1016/j.jml.2011.09.004
110. Shukla M, Nespor M, Mehler J. An interaction between prosody and statistics in the segmentation of fluent speech. Cogn Psychol. 2007;54(1):1–32.
111. Kim S, Broersma M, Cho T. The use of prosodic cues in learning new words in an unfamiliar language. Stud Second Lang Acquis. 2012;34(3):415–44.
112. White L, Benavides-Varela S, Mády K. Are initial-consonant lengthening and final-vowel lengthening both universal word segmentation cues? J Phon [Internet]. 2020;81:100982. Available from: https://doi.org/10.1016/j.wocn.2020.100982
113. Morgan JL, Saffran JR. Emerging Integration of Sequential and Suprasegmental Information in Preverbal Speech Segmentation. Child Dev. 1995;66(4):911–36.
114. Trainor LJ, Adams B. Infants ' and adults ' use of duration and intensity cues in the segmentation of tone patterns. Percept Psychophys. 2000;62(2):333–40.
115. Zellner B. Pauses and the Temporal Structure of Speech. In: Keller E, editor. Fundamentals of speech synthesis and speech recognition. Chichester: John Wiley; 1994. p. 41–62.
116. Matzinger T, Ritt N, Fitch WT. Non-native speaker pause patterns closely correspond to those of native speakers at different speech rates. PLoS One [Internet]. 2020;15(4):1–20. Available from: http://dx.doi.org/10.1371/journal.pone.0230710
117. Derwing TM, Munro MJ, Thomson RI, Rossiter MJ. The relationship between L1 fluency and L2 fluency development. Stud Second Lang Acquis. 2009;31(4):533–57.
118. Kershenbaum A, Blumstein DT, Roch MA, Akçay Ç, Bee MA, Bohn K, et al. Acoustic sequences in non-human animals: a tutorial review and prospectus. Biol Rev Camb Philos Soc. 2016;91(1):13–52.
119. Mann DC, Hoeschele M. Segmental units in nonhuman animal vocalization as a window into meaning, structure, and the evolution of language. Anim Behav Cogn. 2020;7(2):151–8.
120. Klatt DH. Linguistic uses of segmental duration in English: acoustic and perceptual evidence. J Acoust Soc Am. 1975;59(5):1208–20.
121. Oller DK. The effect of position in utterance on speech segment duration in English. J Acoust Soc Am. 1973;54(5):1235–47.
122. Vaissière J. Language-independent prosodic features. In: Cutler A, Ladd DR, editors. Springer Series in Language and Communication 14: Prosody: Models and Measurements. Berlin: Springer; 1983. p. 53–66.
123. Seifart F, Strunk J, Danielsen S, Hartmann I, Pakendorf B, Wichmann S, et al. The extent and degree of utterance-final word lengthening in spontaneous speech from 10 languages. Linguist Vanguard. 2021;7(1):1–14.
124. Berkovits R. Durational effects in final lengthening, gapping, and contrastive stress. Lang Speech [Internet]. 1994;37(3):237–50. Available from: http://www.ncbi.nlm.nih.gov/pubmed/7861912
125. Byrd D. Articulatory vowel lengthening and coordination at phrasal junctures. Phonetica [Internet]. 2000;57(1):3–16. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10867568
126. Edwards J, Beckman ME, Fletcher J. The articulatory kinematics of final lengthening. J Acoust Soc Am [Internet]. 1991;89(1):369–82. Available from: http://asa.scitation.org/doi/10.1121/1.400674

127. Myers S, Hansen BB. The Origin of Vowel Length Neutralization in Final Position: Evidence from Finnish speakers. Nat Lang Linguist Theory. 2007;25(1):157–93.
128. Krivokapic J. Gestural coordination at prosodic boundaries and its role for prosodic structure and speech planning processes. Philos Trans R Soc B. 2014;369(20130397).
129. Friberg A, Sundberg J. Does music performance allude to locomotion ? A model of final ritardandi derived from measurements of stopping runners. J Acoust Soc Am. 1999;105(3):1469–84.
130. Scott DR. Duration as a cue to the perception of a phrase boundary. J Acoust Soc Am [Internet]. 1982;71(4):996–1007. Available from: http://www.ncbi.nlm.nih.gov/pubmed/7085988
131. Kachkovskaia T. Phrase-final Lengthening in Russian: Pre-boundayr or Pre-pausal? In: Proceedings of the 16th International Conference on Speech and Computer, SPECOM 2014. 2014. p. 353–9.
132. Christiansen MH, Kirby S. Language evolution: Consensus and controversies. Trends Cogn Sci. 2003;7(7):300–7.
133. Bolton TL. Rhythm. Am J Psychol. 1894;6(2):145–238.
134. Woodrow H. A quantitative study of rhythm: The effect of variations in intensity, rate and duration. Arch Psychol. 1909;(14):1–66.
135. Hayes B. Metrical stress theory: Principles and case studies. Chicago: The University of Chicago Press; 1995.
136. Hay JSF, Diehl RL. Perception of rhythmic grouping: Testing the iambic/trochaic law. Pereception Psychophys. 2007;69(1):113–22.
137. De la Mora DM, Nespor M, Toro JM. Do humans and nonhuman animals share the grouping principles of the iambic – trochaic law? Atten Percept Psychophys. 2013;75(1):92–100.
138. Iversen JR, Patel AD, Ohgushi K. Perception of rhythmic grouping depends on auditory experience. J Acoust Soc Am. 2008;124(4):2263–71.
139. Crowhurst M. Iambic-Trochaic Law Effects among Native Speakers of Spanish and English. Lab Phonol. 2016;7(1):12.
140. Crowhurst M, Teodocio Olivares A. Beyond the iambic-trochaic law: The joint influence of duration and intensity on the perception of rhythmic speech. Vol. 31, Phonology. 2014. 51–94 p.
141. Peña M, Bion RAH, Nespor M. How Modality Specific Is the Iambic-Trochaic Law? Evidence From Vision. J Exp Psychol Learn Mem Cogn. 2011;37(5):1199–208.
142. Mann DC, Fitch WT, Tu HW, Hoeschele M. Universal principles underlying segmental structures in parrot song and human speech. Sci Rep [Internet]. 2021;11(1):1–14. Available from: https://doi.org/10.1038/s41598-020-80340-y
143. Tierney AT, Russo FA, Patel AD. The motor origins of human and avian song structure. Proc Natl Acad Sci [Internet]. 2011;108(37):15510–5. Available from: http://www.pnas.org/content/108/37/15510
144. Pika S, Wilkinson R, Kendrick KH, Vernes SC. Taking turns: Bridging the gap between human and animal communication. Proc R Soc B Biol Sci. 2018;285(1880).
145. Ravignani A, Verga L, Greenfield MD. Interactive rhythms across species: the evolutionary biology of animal chorusing and turn-taking. Ann N Y Acad Sci. 2019;1453(1):12–21.
146. Bagou O, Fougeron C, Frauenfelder UH. Contribution of Prosody to the Segmentation and Storage of " Words " in the Acquisition of a New Mini-Language. Proc Speech Prosody 2002 Conf. 2002;(January 2002):159–62.
147. Gordon M, Roettger T. Acoustic correlates of word stress: A cross-linguistic survey. Linguist Vanguard. 2017;3(1):1–11.
148. Andersson S, Sayeed O, Vaux B. The Phonology of Language Contact. In: Oxford

Handbooks Online. 2017. p. 1–33.

149. Broselow E. Stress adaptation in loanword phonology : In: Boersma P, Hamann S, editors. Phonology in perception. Berlin: De Gruyter Mouton; 2009. p. 191–234.

150. Speyer A. On the change of word stress in the history of German. Beitrage zur Geschichte der Dtsch Spr und Lit. 2009;131(3):413–41.

151. Nolan F. Intonation. In: Aarts B, McMahon AMS, Hinrichs L, editors. The Handbook of English Linguistics. John Wiley & Sons; 2021. p. 385–405.

152. Hirst D, Di Cristo A, editors. Intonation systems - a survey of twenty languages. Intonation systems: A survey of twenty languages. Cambridge: Cambridge University Press; 1998.

153. Ellis NC. Frequency effects in language processing. Stud Second Lang Acquis. 2002;24:143–88.

154. Diessel H. Frequency effects in language acquisition, language use, and diachronic change. New Ideas Psychol. 2007;25:108–27.

155. Ambridge B, Kidd E, Rowland CF, Theakston AL. The ubiquity of frequency effects in first language acquisition. J Child Lang. 2015;42:239–73.

156. Toro JM, Sebastián-Gallés N, Mattys SL. The role of perceptual salience during the segmentation of connected speech. Eur J Cogn Psychol. 2009;21(5):786–800.

157. Abboub N, Boll-Avetisyan N, Bhatara A, Höhle B, Nazzi T. An exploration of rhythmic grouping of speech sequences by French- and German-learning infants. Front Hum Neurosci. 2016;10.

158. Bion RAH, Benavides-Varela S, Nespor M. Acoustic markers of prominence influence infants' and adults' segmentation of speech sequences. Lang Speech. 2011;54(1):123–40.

159. Nespor M, Shukla M, Van De Vijver R, Avesani C, Schraudolf H, Donati C. Different phrasal prominence realizations in VO and OV languages. Lingue e Linguaggio. 2008;7(2):139–67.

160. Hauser MD. A Primate Dictionary? Decoding the Function and Meaning of Another Species' Vocalizations. Cogn Sci [Internet]. 2000;24(3):445–75. Available from: http://onlinelibrary.wiley.com/doi/10.1207/s15516709cog2403_5/abstract

161. Pisanski K, Oleszkiewicz A, Plachetka J, Gmiterek M, Reby D. Voice pitch modulation in human mate choice. Proc R Soc B. 2018;285(20181634):1–8.

CHAPTER 2

The influence of different prosodic cues on word segmentation

This chapter is published in *Frontiers in Psychology*.

Citation:

Matzinger T, Ritt N & Fitch WT. 2021. The influence of different prosodic cues on word segmentation. *Frontiers in Psychology* 12: 622042. https://doi.org/10.3389/fpsyg.2021.622042

frontiers
in Psychology

ORIGINAL RESEARCH
published: 16 March 2021
doi: 10.3389/fpsyg.2021.622042

The Influence of Different Prosodic Cues on Word Segmentation

Theresa Matzinger[1,2]*, *Nikolaus Ritt*[1] *and W. Tecumseh Fitch*[2,3]*

[1] Department of English, University of Vienna, Vienna, Austria, [2] Department of Behavioral and Cognitive Biology, University of Vienna, Vienna, Austria, [3] Cognitive Science Hub, University of Vienna, Vienna, Austria

A prerequisite for spoken language learning is segmenting continuous speech into words. Amongst many possible cues to identify word boundaries, listeners can use both transitional probabilities between syllables and various prosodic cues. However, the relative importance of these cues remains unclear, and previous experiments have not directly compared the effects of contrasting multiple prosodic cues. We used artificial language learning experiments, where native German speaking participants extracted meaningless trisyllabic "words" from a continuous speech stream, to evaluate these factors. We compared a baseline condition (statistical cues only) to five test conditions, in which word-final syllables were either (a) followed by a pause, (b) lengthened, (c) shortened, (d) changed to a lower pitch, or (e) changed to a higher pitch. To evaluate robustness and generality we used three tasks varying in difficulty. Overall, pauses and final lengthening were perceived as converging with the statistical cues and facilitated speech segmentation, with pauses helping most. Final-syllable shortening hindered baseline speech segmentation, indicating that when cues conflict, prosodic cues can override statistical cues. Surprisingly, pitch cues had little effect, suggesting that duration may be more relevant for speech segmentation than pitch in our study context. We discuss our findings with regard to the contribution to speech segmentation of language-universal boundary cues vs. language-specific stress patterns.

Keywords: language learning, speech segmentation, prosody, statistical cues, word stress, pauses

OPEN ACCESS

Edited by:
Juhani Järvikivi,
University of Alberta, Canada

Reviewed by:
Carolyn Quam,
Portland State University,
United States
Isabell Wartenburger,
University of Potsdam, Germany

****Correspondence:***
Theresa Matzinger
theresa.matzinger@univie.ac.at
W. Tecumseh Fitch
tecumseh.fitch@univie.ac.at

Specialty section:
This article was submitted to
Language Sciences,
a section of the journal
Frontiers in Psychology

Received: *27 October 2020*
Accepted: *02 February 2021*
Published: *16 March 2021*

Citation:
Matzinger T, Ritt N and Fitch WT
(2021) The Influence of Different
Prosodic Cues on Word
Segmentation.
Front. Psychol. 12:622042.
doi: 10.3389/fpsyg.2021.622042

INTRODUCTION

The Speech Segmentation Problem

When people begin acquiring a new language, a particular challenge is the segmentation of fluent speech into words. This task is especially difficult because continuous speech lacks directly accessible cues to word boundaries. Prominent acoustic cues, such as pauses, are rare and occur only inconsistently (Cole et al., 1980; Saffran et al., 1996a; Cutler et al., 1997; Johnson, 2008). This initial speech segmentation problem is most acute for infants learning their first language but is also daunting for second language learners. For adults, the challenge is particularly apparent when they try to identify discrete words in an unfamiliar foreign language (Johnson and Jusczyk, 2001; Endress and Hauser, 2010; Erickson and Thiessen, 2015). Nonetheless, language learners eventually master the speech segmentation problem with ease.

Experimental Paradigm and Study Rationale

The mechanisms and cues that potentially help language learners extract words from continuous speech have been the subject of a large body of previous research on both infants and adults (e.g., Saffran et al., 1996a,b, 1999; Aslin et al., 1998; Johnson and Jusczyk, 2001; Johnson, 2008, 2012; Johnson and Seidl, 2009; Tyler and Cutler, 2009; Johnson and Tyler, 2010; Hay and Saffran, 2012; Frost et al., 2017). Most of this research used the well-established "artificial language learning" paradigm (Saffran et al., 1996a), which models natural language learning. In this paradigm, listeners are exposed for several minutes to a continuous speech stream of nonsense speech, generated by concatenating invented trisyllabic pseudo-words in a random order. Participants are subsequently tested on the recognition of the intended pseudo-words, as opposed to "part-words": syllable sequences that occurred due to the juxtaposition of two pseudo-words, which have lower transitional probabilities. For example, listeners might hear the nonsense speech stream *...bakupodelarufumesigonitedelarubakupogonitefumesi...* and infer the recurring trisyllables *bakupo*, *delaru*, *fumesi* and *gonite* as acceptable pseudo-words, while rejecting the part-words *kupode*, *podela* or similar items because these syllables occur in sequence less frequently (e.g., Saffran et al., 1996a; **Figure 1**: 1. Baseline condition). We will refer to these transitional probabilities between syllables as "statistical cues" and the "words," i.e., the group of three syllables with the highest internal transitions probabilities (*bakupo*, *delaru*, *fumesi*, and *gonite*, in **Figure 1**) as "statistical words" hereafter.

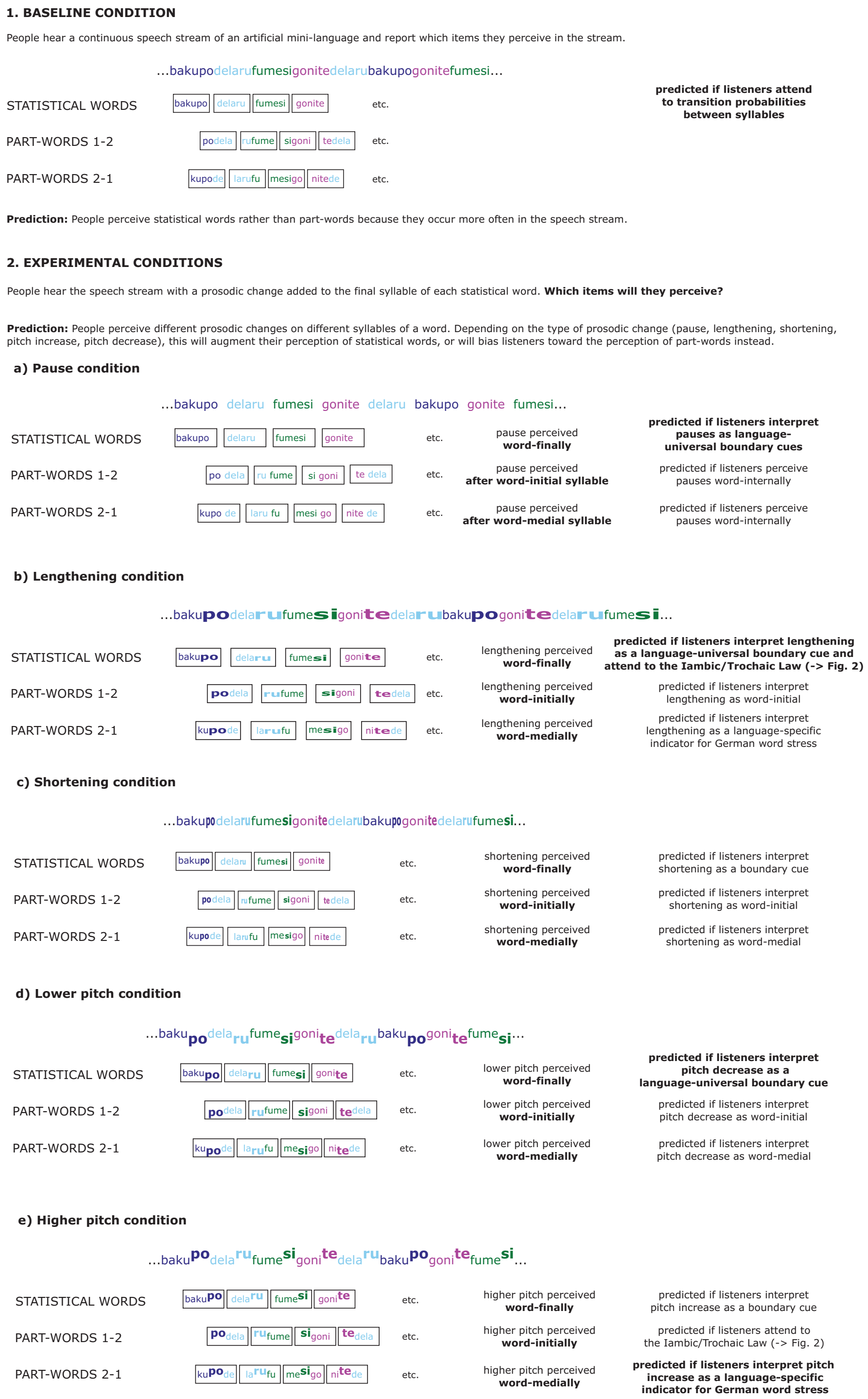
In this study, we adopted the general design above, but added additional acoustic cues to the nonsense speech stream to investigate how such changes influence listeners' speech segmentation. For simplicity, our study focused on the investigation of prosodic cues on word-final syllables. Thus, durational changes and pitch changes were always implemented on the final syllable of the trisyllabic statistical pseudo-words. Our main aim was to investigate how various prosodic cues such as pauses between statistical words, word-final lengthening, word-final shortening, word-final pitch decrease, and word-final pitch increase influenced which three-syllable groupings German speaking participants segmented from the speech stream as "words" (**Figure 1**: 2. Experimental conditions). Our second aim was to test how potential language-universal cognitive predispositions and/or language-specific word stress patterns typical of the listeners' native languages influence speech segmentation in an experimental setting (Tyler and Cutler, 2009; Frost et al., 2017; Ordin et al., 2017). We tested German speaking participants because German word stress patterns (most trisyllabic German words are stressed on word-*medial* syllables; Féry, 1998) contrast nicely with language-universal prosodic boundary cues on word-*final* syllables (e.g., phrase-final lengthening; e.g., Fletcher, 2010). If listeners attend to language-universal prosodic boundary cues, adding such cues to the last syllable of a three-syllable statistical word should be perceived as converging with the statistical cues and therefore should facilitate participants' speech segmentation performance ("cue convergence"). In contrast, if listeners interpret such cues as German stress cues, i.e., if they interpret them as occurring word-medially, the prosodic cues would indicate different word boundaries than the boundaries indicated by the statistical cues. Therefore, in this scenario, adding such prosodic cues to the last syllable of a three-syllable statistical word should be perceived as conflicting with the statistical cues. In this case, prosodic cues would hinder speech segmentation based on statistical cues, or even lead to different segmentation patterns than those expected from attending to transition probabilities alone ("cue conflict"). Thus, our paradigm not only compared different prosodic cues, but also helps to disentangle whether adult participants tend to use language-universal or language-specific prosodic cues during speech segmentation. We will explain the study background and our hypotheses in more detail below; see **Figure 1** for an overview.

Speech Segmentation Strategies and Cue Types

Previous research provided abundant evidence that language learners can draw on multiple sources of information for word segmentation (Johnson and Jusczyk, 2001; Mattys et al., 2005; Filippi et al., 2014; Mitchel and Weiss, 2014; Morrill et al., 2015; Johnson, 2016; Sohail and Johnson, 2016), among which "statistical cues" (i.e., transitional probabilities between syllables) and prosodic cues are very prominent.

Using statistical cues present in the speech stream is a very basic language-universal speech segmentation strategy. This strategy is based on tracking transitional probabilities between syllables, which represent the statistical likelihood that one syllable directly follows another (e.g., Saffran et al., 1996a; Aslin et al., 1998; Romberg and Saffran, 2010; Johnson, 2016). Syllables that co-occur frequently are likely to belong to the same word, whereas syllables that co-occur rarely usually span word boundaries (Hayes and Clark, 1970; Swingley, 2005; Johnson and Seidl, 2009; Hay and Saffran, 2012). For example, in the sound sequence "*principal component,*" the transitional probabilities from *prin* to *ci* to *pal* are higher than from *pal* to *com* because *prin*, *ci* and *pal* also co-occur in other sequences including the word *principal*, such as *principal investigator*, *principal purpose* or *principal reasons*, whereas *pal* and *com* are only rarely found in immediate succession (frequencies in the Corpus of Contemporary American English: *prin-ci*: 114,277 occurrences, *ci-pal*: 57,520 occurrences, *pal-com*: 1,065 occurrences; Davies 2008). Cross-linguistically, listeners are able to track these statistical relationships, and use them to infer which sound sequences constitute words (Saffran et al., 1996a; Aslin et al., 1998). Still, considerable evidence suggests that statistical cues, while powerful, are not the only information that listeners use to segment speech into words (Morgan and Saffran, 1995; Johnson and Jusczyk, 2001; but also: Thiessen and Saffran, 2003; Johnson and Seidl, 2009; Endress and Hauser, 2010; Johnson and Tyler, 2010; Johnson et al., 2014).

Prosodic cues linked to word stress or word boundaries can provide important additions to statistical cues, and typically enhance speech segmentation performance in infants (e.g., Morgan and Saffran, 1995; Mattys et al., 1999; Johnson and Jusczyk, 2001; Thiessen and Saffran, 2003; Seidl, 2007; Johnson and Seidl, 2009) and adults (e.g., Cutler, 1991; Saffran et al., 1996b; Endress and Mehler, 2009; Endress and Hauser, 2010; Frost et al., 2017). Furthermore, phrasal prosody (e.g., Christophe et al., 2004; Gout et al., 2004; Shukla et al., 2007; Mueller et al.,



1. BASELINE CONDITION

People hear a continuous speech stream of an artificial mini-language and report which items they perceive in the stream.

...bakupodelarufumesigonitedelarubakupogonitefumesi...

STATISTICAL WORDS	bakupo	delaru	fumesi	gonite	etc.	**predicted if listeners attend to transition probabilities between syllables**
PART-WORDS 1-2	podela	rufume	sigoni	tedela	etc.	
PART-WORDS 2-1	kupode	larufu	mesigo	nitede	etc.	

Prediction: People perceive statistical words rather than part-words because they occur more often in the speech stream.

2. EXPERIMENTAL CONDITIONS

People hear the speech stream with a prosodic change added to the final syllable of each statistical word. **Which items will they perceive?**

Prediction: People perceive different prosodic changes on different syllables of a word. Depending on the type of prosodic change (pause, lengthening, shortening, pitch increase, pitch decrease), this will augment their perception of statistical words, or will bias listeners toward the perception of part-words instead.

a) Pause condition

...bakupo delaru fumesi gonite delaru bakupo gonite fumesi...

STATISTICAL WORDS	bakupo	delaru	fumesi	gonite	etc.	pause perceived **word-finally**	**predicted if listeners interpret pauses as language-universal boundary cues**
PART-WORDS 1-2	po dela	ru fume	si goni	te dela	etc.	pause perceived **after word-initial syllable**	predicted if listeners perceive pauses word-internally
PART-WORDS 2-1	kupo de	laru fu	mesi go	nite de	etc.	pause perceived **after word-medial syllable**	predicted if listeners perceive pauses word-internally

b) Lengthening condition

...bakupodelarufumesigonitedelarubakupogonitedelarufumesi...

STATISTICAL WORDS	bakupo	delaru	fumesi	gonite	etc.	lengthening perceived **word-finally**	**predicted if listeners interpret lengthening as a language-universal boundary cue and attend to the Iambic/Trochaic Law (-> Fig. 2)**
PART-WORDS 1-2	podela	rufume	sigoni	tedela	etc.	lengthening perceived **word-initially**	predicted if listeners interpret lengthening as word-initial
PART-WORDS 2-1	kupode	larufu	mesigo	nitede	etc.	lengthening perceived **word-medially**	predicted if listeners interpret lengthening as a language-specific indicator for German word stress

c) Shortening condition

...bakupodelarufumesigonitedelarubakupogonitedelarufumesi...

STATISTICAL WORDS	bakupo	delaru	fumesi	gonite	etc.	shortening perceived **word-finally**	predicted if listeners interpret shortening as a boundary cue
PART-WORDS 1-2	podela	rufume	sigoni	tedela	etc.	shortening perceived **word-initially**	predicted if listeners interpret shortening as word-initial
PART-WORDS 2-1	kupode	larufu	mesigo	nitede	etc.	shortening perceived **word-medially**	predicted if listeners interpret shortening as word-medial

d) Lower pitch condition

...bakupodelarufumesigonitedelarubakupogonitefumesi...

STATISTICAL WORDS	bakupo	delaru	fumesi	gonite	etc.	lower pitch perceived **word-finally**	**predicted if listeners interpret pitch decrease as a language-universal boundary cue**
PART-WORDS 1-2	podela	rufume	sigoni	tedela	etc.	lower pitch perceived **word-initially**	predicted if listeners interpret pitch decrease as word-initial
PART-WORDS 2-1	kupode	larufu	mesigo	nitede	etc.	lower pitch perceived **word-medially**	predicted if listeners interpret pitch decrease as word-medial

e) Higher pitch condition

...bakupodelarufumesigonitedelarubakupogonitefumesi...

STATISTICAL WORDS	bakupo	delaru	fumesi	gonite	etc.	higher pitch perceived **word-finally**	predicted if listeners interpret pitch increase as a boundary cue
PART-WORDS 1-2	podela	rufume	sigoni	tedela	etc.	higher pitch perceived **word-initially**	predicted if listeners attend to the Iambic/Trochaic Law (-> Fig. 2)
PART-WORDS 2-1	kupode	larufu	mesigo	nitede	etc.	higher pitch perceived **word-medially**	**predicted if listeners interpret pitch increase as a language-specific indicator for German word stress**

FIGURE 1 | Overview of the study design and predictions. “Part-words 1–2” are created from the final syllable of a statistical word and the initial and medial syllable of the following statistical word. “Part-words 2–1” are created from the medial and final syllable of a statistical word and the initial syllable of the following statistical word. If prosodic cues converge with the statistical cues, participants will perceive the “statistical words.” If prosodic cues conflict with the statistical cues, participants will be biased toward perceiving part-words. The right column contains the most important predictions. Predictions that are derived from previous studies and are therefore most likely to be borne out (for a more detailed discussion, see main text) are highlighted in bold. Predictions that are not informed by evidence-based language-universal and language-specific considerations or are less likely to be borne out are displayed in normal font.

2010; Langus et al., 2012) and speech pauses (e.g., Johnson et al., 2014; Sohail and Johnson, 2016) facilitate speech segmentation. Compared to statistical cues, which require computations over large sets of syllables, prosodic cues can be extracted relatively directly from the immediate acoustic stimulus (Christophe et al., 2004; Gout et al., 2004; Johnson and Seidl, 2009; Hay and Saffran, 2012; Erickson and Thiessen, 2015), making it reasonable that language learners, especially infants, use them to help solve the speech segmentation problem.

Crucially, prosodic cues can manifest in multiple independent acoustic correlates such as changes in syllable duration, pitch, or loudness, and different acoustic correlates can have different separable effects on speech segmentation (Hay and Saffran, 2012; Ordin and Nespor, 2013). Many previous studies used a combination of different acoustic correlates, but did not determine which prosodic cues were most relevant for word segmentation (Johnson and Jusczyk, 2001; Thiessen and Saffran, 2003, 2007; Johnson and Seidl, 2009). Multiple studies have examined the role of individual cues, suggesting that lengthening serves as a language-universal signal for word-finality (Tyler and Cutler, 2009; Hay and Saffran, 2012; Kim et al., 2012; Frost et al., 2017; but also: White et al., 2020), and that pitch increase is a signal for word stress and is therefore processed differently by speakers of different languages (Morgan and Saffran, 1995; for infants see e.g., Johnson and Jusczyk, 2001; Tyler and Cutler, 2009; Ordin and Nespor, 2016). However, direct comparisons of the roles of different prosodic cues for word segmentation are scarce (e.g., Tyler and Cutler, 2009). To our knowledge, cue changes that contrast in their direction (such as lengthening vs. shortening, or pitch increase vs. decrease) have not been investigated in direct comparison before.

Also, in artificial language learning experiments, prosodic cues that are linked to word stress or word boundaries should only facilitate speech segmentation compared to a statistical baseline if listeners perceive the prosodic cues as converging with the statistical cues defined by the transition probabilities between syllables in the speech stream. For example, in our experiment, if listeners interpret lengthening as a signal for word-finality and perceive it as occurring in word-final position, lengthening should facilitate speech segmentation, since in our experiment, lengthening was always implemented on the final syllable of statistical words. In contrast, if listeners interpret lengthening as a signal for word-initial or word-medial position, listeners should interpret lengthening of the last syllable of statistical words in our experiments as a conflicting cue. In such a case, where prosodic cues conflict with the available statistical cues, prosodic cues could potentially impair speech segmentation relative to statistical cues alone, or even override them and lead to different segmentation patterns (Johnson and Jusczyk, 2001; Thiessen and Saffran, 2003; Johnson and Seidl, 2009; Ordin and Nespor, 2013). Hereafter, we will follow the convention of previous speech segmentation studies (e.g., Frost et al., 2017; Ordin et al., 2017) by defining segmentation that is based on statistical words (potentially enhanced by converging prosodic cues) as the "correct" segmentation. In contrast, if listeners based their segmentation decisions on prosodic cues that conflict with statistical cues, e.g., because they applied a segmentation strategy based on German-specific word-stress patterns, this will be defined in our analyses as "impaired" or "incorrect" segmentation relative to the statistical word baseline. Obviously, such segmentation strategies can also lead to the consistent extraction of items from the speech stream in experimental settings, and there is no intrinsic right or wrong answer in experiments using pseudo-words, but the items resulting from such segmentation strategies clearly differ from the words based on statistical cues alone, which we will henceforth term "correct." Also, note that our use of statistical cues as a baseline is a product of our experimental design and analysis, and we use the terms "baseline" and "correct" for convenience in describing our results. We clearly do not intend to suggest that statistical cues are somehow primary or "correct" in real-world speech segmentation (and indeed we suspect that prosodic cues might often be dominant): the relative strength of these factors is precisely what our experiments set out to test.

Choice of Prosodic Cues in Our Study

Although speech segmentation has been widely investigated, it remains unclear which specific acoustic correlates of prosody, such as changes in syllable duration or pitch, are most relevant for speech segmentation. Therefore, the main aim of the current study was to investigate the relative contribution that different acoustic manifestations of prosody make toward speech segmentation in adults. Our study focused on five different prosodic changes in three different acoustic cue categories (**Figure 1**: 2. Experimental conditions). These were durational cues: (a) syllable lengthening and (b) syllable shortening; voice fundamental frequency or "pitch" cues: (c) pitch increase and (d) pitch decrease; and (e) pause cues (intervals of silence between statistical words). We compared these five individual prosodic changes to a baseline condition that included only statistical cues, i.e., transition probabilities between syllables (**Figure 1**). This comparison of multiple word segmentation cues, including contrasting prosodic cues, within a single study sets our study apart from previous speech segmentation studies.

We chose these five cues because pauses, durational, and pitch cues can function either as language-universal cues to word boundaries or as language-specific cues to word stress. Some of the cues have been shown to signal boundaries and word stress more successfully than others. Lengthening and pitch increase have been previously investigated in similar contexts (e.g., Saffran et al., 1996b; Frost et al., 2017; Ordin et al., 2017), but rarely in direct comparison (as in Tyler and Cutler, 2009). Most likely, past studies have focused on lengthening and pitch increase because both of these cues are typical acoustic correlates for expressing language-specific word-stress (Tyler and Cutler, 2009), and final lengthening is a cross-linguistic signal for word, phrase and sentence boundaries (Fletcher, 2010). Interestingly, shortened duration and decreased pitch have been neglected in past research on word segmentation (but see research on pitch decrease in a phrasal context; Mueller et al., 2010), presumably because these changes normally do not signal word stress. Still, they may provide valuable comparisons to lengthening and pitch increase to see if prosodic patterns that are not typical word stress correlates, and may even contrast with typical word stress correlates in natural languages, can still facilitate speech segmentation in an experimental setting. Further, the manipulation of acoustic cues that are not typical stress correlates of target words may lead to insights about how these cues may influence speech segmentation when occurring in a more distal prosodic context in real-life speech processing (cf. Dilley and McAuley, 2008).

Besides durational and pitch cues, intensity is a typical acoustic correlate of stress. We did not include intensity in our study because its role as a perceptual correlate of stress is unclear and because intensity levels are usually correlated with vowel quality and duration (Cutler, 2005; Ordin and Nespor, 2013).

Pauses, our third cue category, represent a language-universal boundary cue that should be salient independent of listeners' preferred stress patterns since they do not serve to signal word stress (Fletcher, 2010; Johnson, 2016). Pauses thus serve as reference cues for segmentation (Peña et al., 2002). Also, pauses are interesting because speech input consisting of words separated by pauses may help infant word learning less than continuous speech (Johnson et al., 2013). Crucially, pauses have a durational component and can be longer or shorter, but we regard them as a separate cue category because they differ from our syllable durational cues (lengthening and shortening) in many other aspects. For example, silent pauses do not consist of any acoustic material and thus cannot signal word stress.

We chose to focus on modifications of word-*final* syllables because in natural languages, final elements are often particularly susceptible to modifications (Swingley, 2009), e.g., in phrase-final lengthening (Fletcher, 2010), reduction of word-final unstressed syllables (Kohler and Rodgers, 2001; O'Brien and Fagan, 2016), or utterance-final pitch lowering in declarative sentences (Cruttenden, 1986; Hirst and Di Cristo, 1998). Also, pitch changes and durational changes implemented *on* word-final syllables can easily be compared to pause cues between words (that is, *after* word-final syllables). Modifying word-final syllables is also interesting insofar as this contrasts nicely with the dominant word stress pattern of our participants' native language, German, which carries stress predominantly on medial syllables of trisyllabic words (see below; Féry, 1998). If participants interpret the modified word-final syllables in our experiment as being stressed and relate this to the typical word-stress patterns of German, they may interpret the modifications to occur word-medially. This is particularly plausible for typical stress correlates such as pitch increase and lengthening, and may lead to a potential conflict between statistical cues (i.e., transition probabilities between syllables in the experimental speech stream) and prosodic cues. Such an effect would help to evaluate the relative influence of language-specific stress patterns and language-universal boundary cues on speech segmentation (cf. Crowhurst, 2016). Although it would certainly be interesting to test stress cues in other positions as well (Saffran et al., 1996b; Tyler and Cutler, 2009; Ordin et al., 2017; cf. Frost et al., 2017), the large number of acoustic cues we manipulated did not allow us to also investigate word-initial and word-medial changes.

Word Stress in German

We focused on German, a stress-based language (Pamies Bertrán, 1999) that suits itself to theoretically grounded predictions, but is relatively underrepresented in speech segmentation research. Fortunately, a few speech segmentation studies on German (e.g., Bhatara et al., 2013; Ordin and Nespor, 2016; Ordin et al., 2017; Marimon Tarter, 2019), were available to inform our predictions and stimulus choice. In German, word stress in trisyllabic words is variable and depends on syllable structure (for in-depth discussions, see e.g., Delattre, 1965; Giegerich, 1985; Féry, 1998; Dogil and Williams, 1999; Domahs et al., 2014). Still, crucially, about half of all German trisyllabic words are stressed on their medial syllable, and word-initial or word-final stress occur less frequently (Féry, 1998). Similar relations hold for the syllable structures used in our study (see methods section; Féry, 1998; Ernestus and Neijt, 2008; Domahs et al., 2014). Thus, to the extent that listeners are sensitive to statistical regularities in speech, they should assume word-medial stress as the default German stress pattern when encountering new lexical items. If the stress pattern of our listeners' native language affects cue perception, this predicts that stress cues implemented on medial syllables of trisyllabic words should be perceived as converging with statistical cues (transitional probabilities between syllables), whereas stress cues implemented on word-initial or word-final syllables should be less convergent and may even conflict with statistical cues. Thus, German stress patterns contrast nicely with proposed language-universal cues such as phrase-final or sentence-final lengthening (e.g., Fletcher, 2010). If native German speaking listeners attend to a language-universal final lengthening cue, rather than to their dominant native stress pattern, our listeners should perceive word-final lengthening as a cue that strongly converges with the statistical cues, i.e., the transitional probabilities in the speech stream.

In German speech, stressed syllables are both longer and higher pitched than unstressed syllables (Ordin et al., 2017), but evidence about which of these two manifestations plays a bigger role for production and perception is inconclusive (pitch: Isachenko and Schädlich, 1966; syllable duration: Dogil and Williams, 1999; Nespor et al., 2008; Féry et al., 2011; Kohler, 2012; El Zarka et al., 2017). There are previous indications that in German, lengthening cues are perceived as converging with statistical cues when they occur in word-final position (Ordin and Nespor, 2013, 2016; Ordin et al., 2017), possibly because the cross-linguistic tendency to lengthen word final syllables (e.g., Fletcher, 2010) overrides the perception of the typical German word-medial stress pattern in these cases. Thus, German speakers may focus on pitch as a more reliable cue to word stress instead (cf. Kohler, 2012 on perceptual correlates of stress in German; Nespor et al., 2008; Féry et al., 2011), though this has not been observed experimentally (Ordin and Nespor, 2016).

Finally, testing opposing changes, such as lengthening vs. shortening of duration, or increase vs. decrease of pitch, represents a potentially important extension to previous findings on word segmentation in German, where only one direction of change in these cues was tested, because results will show whether *any* arbitrary durational or pitch modification acts as a segmentation cue (e.g., due to difference of any sort), or whether the directionality of the changes is important. To our knowledge, neither opposing cues nor pause cues have previously been tested in word segmentation experiments with German adults. Thus, overall, both theoretical and empirical considerations make German a particularly interesting language for our study.

Hypotheses and Predictions

Our experimental setup given our chosen acoustic parameters leads to several hypotheses and predictions. The first hypothesis is that native German speaking listeners will interpret prosodic cues that occur either on (for durational and pitch cues) or after (for pause cues) the final syllable of statistical words as boundary signals that support the statistical cues already available (cue convergence). This predicts that adding prosodic cues on the word-final syllables will improve listeners' speech segmentation compared to their performance based on statistical cues alone. We refer to this hypothesis, where statistical cues and the individual prosodic cues are perceived as converging, as the "cue convergence hypothesis."

The cue convergence hypothesis can be put forward for each of our prosodic cues separately, though it is more plausible for some changes than for others. Pause cues might be associated with word boundaries because in everyday speech, perceptible pauses occur almost exclusively at word boundaries, and hardly ever within words (Trainor and Adams, 2000; Fletcher, 2010; Sohail and Johnson, 2016; Matzinger et al., 2020). Lengthened syllables might also serve as signals for word-finality because domain-final elements are lengthened in everyday speech language-universally (Oller, 1973; Klatt, 1975; Vaissière, 1983; Tyler and Cutler, 2009; Fletcher, 2010; but also: White et al., 2020). Although domain-final lengthening mostly happens at the sentence or phrase level, we predict that it will generalize to the word level in our study, because in our design each statistical word is essentially a phrase, and there is evidence for successful speech segmentation based on final lengthening cues from previous speech segmentation experiments in several languages, including German (e.g., Saffran et al., 1996b; Tyler and Cutler, 2009; Ordin and Nespor, 2016; Frost et al., 2017; Ordin et al., 2017).

Furthermore, the putatively language-independent Iambic/Trochaic Law (= ITL; Bolton, 1894; Woodrow, 1909; Hayes, 1995; Hay and Diehl, 2007; De la Mora et al., 2013; Frost et al., 2017; but see Iversen et al., 2008) states that listeners group sounds with longer duration as sequence-final (iambic grouping). Although the ITL focuses on disyllabic words, it can also be generalized to trisyllabic words (Trainor and Adams, 2000; Frost et al., 2017), supporting the prediction that final lengthening cues will converge with the available statistical cues and facilitate speech segmentation (**Figure 2**). In contrast, shortened syllables might also potentially signal word boundaries because, in natural languages, word-final elements are frequently phonetically reduced (Kohler and Rodgers, 2001; O'Brien and Fagan, 2016). This is because word processing is incremental and word-final elements are often highly predictable and thus not as informative for word identification as word-initial elements (Dahan and Magnuson, 2006; Swingley, 2009; Wedel et al., 2019).

Pitch decreases may signal word-finality because a sentence-final or phrase-final pitch decrease is very common in natural languages (Vaissière, 1983; Hirst and Di Cristo, 1998; Langus et al., 2012). Again, because in our study design each statistical word equals a phrase, this may generalize to the word level in our study. Finally, word-final pitch *increase* has also been shown to facilitate word segmentation in French, a language with word-final stress (Bagou et al., 2002; Tyler and Cutler, 2009), but not in German adults (Ordin and Nespor, 2016). Thus, overall, all five prosodic changes might potentially converge in word-final position with statistical cues, i.e., transition probabilities, and facilitate word segmentation. However, because of the perceptual salience of pauses, the abundant previous evidence for final lengthening (e.g., Ordin and Nespor, 2016; Ordin et al., 2017) and more tentative evidence against final pitch increase (Ordin and Nespor, 2016) as a speech segmentation cue in German, we predicted the cue convergence hypothesis to apply most strongly for pauses and lengthening, moderately strongly for pitch decrease, and less so for shortening and pitch increase.

An alternative to the cue convergence hypothesis is that native German speaking listeners may interpret prosodic cues implemented on the final syllable of a trisyllabic statistical word as conflicting with the statistical cues provided by the transition probabilities in the speech stream. If participants perceive the modified syllables as being stressed, and then group the syllables in the speech stream in a way that matches the predominant word-medial stress pattern of German (Norris and Cutler, 1988; Cutler, 1990; Cutler et al., 1992; Ordin et al., 2017), the prosodic modifications would then conflict with statistical cues. Since most German trisyllabic words are stressed on the medial syllable, this "cue conflict" hypothesis predicts that placing stress cues on the final syllable of the statistically defined words should bias German listeners' toward a different speech segmentation pattern than that based on statistical cues. Instead, they should group the modified syllables word-medially (see **Figure 1**, Parts 2b, 2d and 2e). We refer to this hypothesis as the "cue conflict hypothesis."

The cue conflict hypothesis is plausible for typical correlates of stress, i.e., lengthening and pitch increase (Thiessen and Saffran, 2003; Johnson and Seidl, 2009), and less so for shortening and pitch decrease. Still, given abundant evidence from previous speech segmentation experiments in several languages, word-final lengthening is expected to converge with the statistical cues, overriding the tendency of native German speaking listeners to interpret lengthening as a cue to word stress (e.g., Ordin and Nespor, 2016; Ordin et al., 2017). Instead, native German speaking listeners are predicted to mostly use pitch increase as a cue for word stress, which would lead to a cue conflict with statistical cues for pitch increase only (contra Ordin and Nespor, 2016). Also, according to the ITL (Hayes, 1995; Nespor et al., 2008; Bion et al., 2011; De la Mora et al., 2013; Abboub et al., 2016), cross-linguistically, listeners group sounds with a higher pitch as sequence-initial (trochaic grouping). Thus, word-final pitch increase might conflict with statistical cues and lead to a different speech segmentation pattern (**Figure 2**). Furthermore, if listeners associate certain prosodic changes with

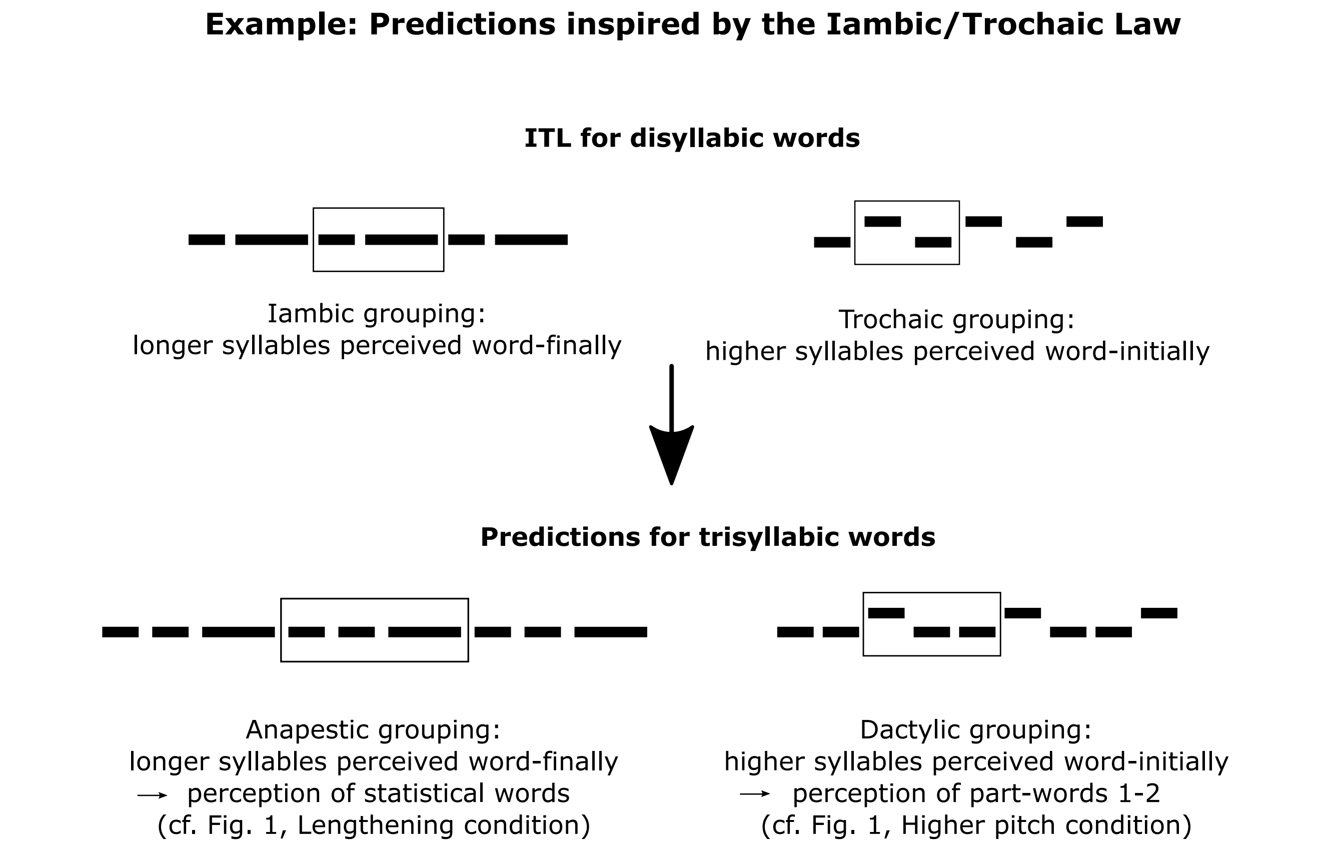

Example: Predictions inspired by the Iambic/Trochaic Law
ITL for disyllabic words
Iambic grouping:
longer syllables perceived word-finally
Trochaic grouping:
higher syllables perceived word-initially
Predictions for trisyllabic words
Anapestic grouping:
longer syllables perceived word-finally
→ perception of statistical words
(cf. Fig. 1, Lengthening condition)
Dactylic grouping:
higher syllables perceived word-initially
→ perception of part-words 1-2
(cf. Fig. 1, Higher pitch condition)

FIGURE 2 | The Iambic/Trochaic Law (ITL) for disyllabic words leads to predictions for how listeners might perceive lengthened and/or higher-pitched syllables in trisyllabic words in our study. Horizontal black bars denote syllables.

word-final syllables (as per the cue convergence hypothesis), they should accordingly associate opposing changes with non-final syllables. Thus, if e.g., lengthening or pitch decrease on the final syllable of statistical words facilitate speech segmentation, the opposing changes (shortening or pitch increase, respectively) can be predicted to lead to a modified segmentation pattern.

In conclusion, for each prosodic cue, both hypotheses might reasonably be expected to hold, but overall, the preponderance of existing evidence suggests that pauses, final lengthening and final pitch decrease will lead to cue convergence, and final shortening and final pitch increase will conflict with statistical cues.

Regarding the relative effects of different prosodic cues, we hypothesized that pauses should have a bigger impact on word segmentation than other prosodic cues. Pauses may provide more salient signals than the other prosodic cues because they involve a highly perceptible decrease in signal amplitude (Fletcher, 2010; Friederici and Männel, 2013). Also, long enough pauses can make a word appear isolated. We thus predicted that word segmentation performance should show a greater increase with pause cues inserted between the "correct" statistical words than for our other prosodic changes. Beyond that basic prediction, durational cues and pitch cues might have different relative strengths, but we had no clear predictions about directionality, given weak and partly inconclusive previous data (cf. Tyler and Cutler, 2009), with some evidence for a durational preference (Männel and Friederici, 2016) and other evidence for a pitch preference (Ordin et al., 2017).

Experimental Variations

Recently, many psychological findings have been found to be non-replicable, commonly known as the replication crisis (Shrout and Rodgers, 2018). Common reasons for a lack of replicability and generalizability are that experimental results are not robust to minor methodological changes (Munafò and Smith, 2018). To counteract this problem in our study, we conducted three experiments that examined whether participants would use similar segmentation strategies when testing paradigm and testing context varied. Our main aim was to evaluate the robustness of our results, and not to pin down effects of specific methodological differences. Therefore, our prime goal was not to design experiments that varied only in a single, carefully controlled methodological feature, but rather to have a spectrum of methods, in a single publication, that roughly mirror the methodological variation typifying previously published speech segmentation studies.

The three experiments implemented the same stimulus manipulations, but differed slightly in experimental setup. Experiment 1 was our initial pilot study, carried out in the participants' normal study or office environment; this study had minimal auditory memory requirements, and combined auditory and visual modalities, i.e., participants could see the test stimuli while they listened to the speech stream. This experiment addressed whether attested laboratory results replicate in an environment where background noise and visual distraction more closely resembled a real-life language learning context. Experiment 3 resembled existing speech segmentation experiments most closely (e.g., Tyler and Cutler, 2009; Frost et al., 2017; Ordin et al., 2017): it was done in a laboratory setting, exclusively in the auditory modality (similar to real-life first language acquisition), and thus involved a strong auditory memory component. However, in contrast to our experiment, where participants decided for single test stimuli if they were statistical words or part-words, most previous adult studies used a two-alternative forced choice testing procedure in which participants had to decide from a set of two test stimuli which of them was a word and which a part-word (see methods for Experiment 3 below). Experiment 2 was designed to be intermediate between Experiments 1 and 3. It was carried out in a laboratory setting, but involved auditory and visual modalities, with minimal memory components. We predicted that the effects of adding prosodic cues to the speech stream might unfold more strongly in experiments with syllables spoken by a native German speaker and a minimal memory component because the overall cognitive load is lower, and statistical cues are less prominent. Also, we expected all effects to be stronger in the laboratory, where people were less distracted than in a natural testing environment (cf. Toro et al., 2005; Erickson and Thiessen, 2015). Nonetheless, if the effects observed are robust and generalizable, they should occur—though perhaps less prominently—both in the natural environment in Experiment 1 because real language learning typically happens in a natural environment, and with an added memory component in Experiment 3 because language learning obviously involves memory (Palmer and Mattys, 2016; Wen, 2016; Pierce et al., 2017).

Additionally, syllables in Experiment 1 were recorded by a native speaker of English, whereas syllables in Experiments 2 and 3 were recorded by a native speaker of German. It is possible that sub-phonemic cues in the native English syllables may influence participants to rely less on their implicit knowledge of German prosody in Experiment 1 than in Experiments 2 and 3 (Quam and Creel, 2017). However, again, if the effects studied in our series of experiments are robust and generalizable, they should also occur in Experiment 1.

GENERAL METHODS

Experimental Paradigm: Overview

We conducted three individual experiments with adult listeners. All three were artificial language learning experiments following an established experimental paradigm (e.g., Saffran et al., 1996a,b, 1999; Frost et al., 2017). Participants in all three experiments listened to a continuous speech stream that was created from four randomly generated trisyllabic pseudo-words making up an artificially constructed pseudo-language, and had to decide for each of 12 test stimuli whether they were "words" of the artificial pseudo-language or not. The study protocol was approved by the ethics board of the University of Vienna (reference number: #00333/00385), and all participants gave written informed consent in accordance with the Declaration of Helsinki.

Experimental Conditions: Overview

In the three experiments, we addressed the influence of different prosodic cues on word segmentation in a baseline and five prosodic conditions, resulting in six conditions in total (see **Figure 3**). In each prosodic condition, the speech stream was manipulated differently to check if that would provide cues to the segmentation of the words from the stream. These changes were always applied after (for pauses) or on (for duration and pitch) the final syllable of each trisyllabic word in the baseline statistical speech stream. Individual syllables were recordings of the same female speaker, but all manipulations of these basic syllables were precisely controlled by computer (for details see "Stimuli," below).

1. Statistical cue only condition (baseline condition). The only cue indicating word segmentation in the baseline condition was that syllable pairs within words had higher transitional probabilities than syllables crossing word-boundaries. Syllables within a word always co-occurred, resulting in within-word transitional probabilities of 1.0. In contrast, each word was pseudo-randomly followed by any of three different words, yielding a between-word syllable transitional probability of 0.33. Thus, participants could potentially infer that syllable pairs that occur more frequently together constitute a word, and those that co-occur less frequently do not. This statistical information was present in all conditions. Each syllable was normalized to a duration of 500 ms and a fundamental frequency of 210 Hz (for details see "Stimuli," below). Typical syllable durations in speech stream experiments conducted in a laboratory are shorter than 500 ms (e.g., Saffran et al., 1996a; Tyler and Cutler, 2009; Frost et al., 2017; Ordin et al., 2017), but since we expected attentional capacities to be limited in Experiment 1, which was conducted in a natural environment, we chose a slow speech rate (2 syllables/second; Song et al., 2010; Palmer and Mattys, 2016) more typical for infant directed speech. This was expected to facilitate speech segmentation.

2. Pause condition. This condition was identical to the baseline condition, with the exception that in addition to the statistical cues, a short pause (250 ms) was inserted after each statistical word. We chose a pause duration of 250 ms because this duration is frequently chosen as a lower detection threshold in studies investigating the occurrence and perception of speech pauses (e.g., Zellner, 1994; Kahng, 2014).

3. Lengthening condition. This condition was identical to the baseline condition, except that in addition to the statistical cues, the final syllable of each word was lengthened by 50%, yielding a duration of 750 ms (cf. previous lengthening by ∼40%: Saffran et al., 1996b; Ordin and Nespor, 2016; Frost et al., 2017; Ordin et al., 2017; or lengthening by 67%: Thiessen and Saffran, 2003). The duration of this additional lengthening was therefore identical to the pause duration in the pause condition.

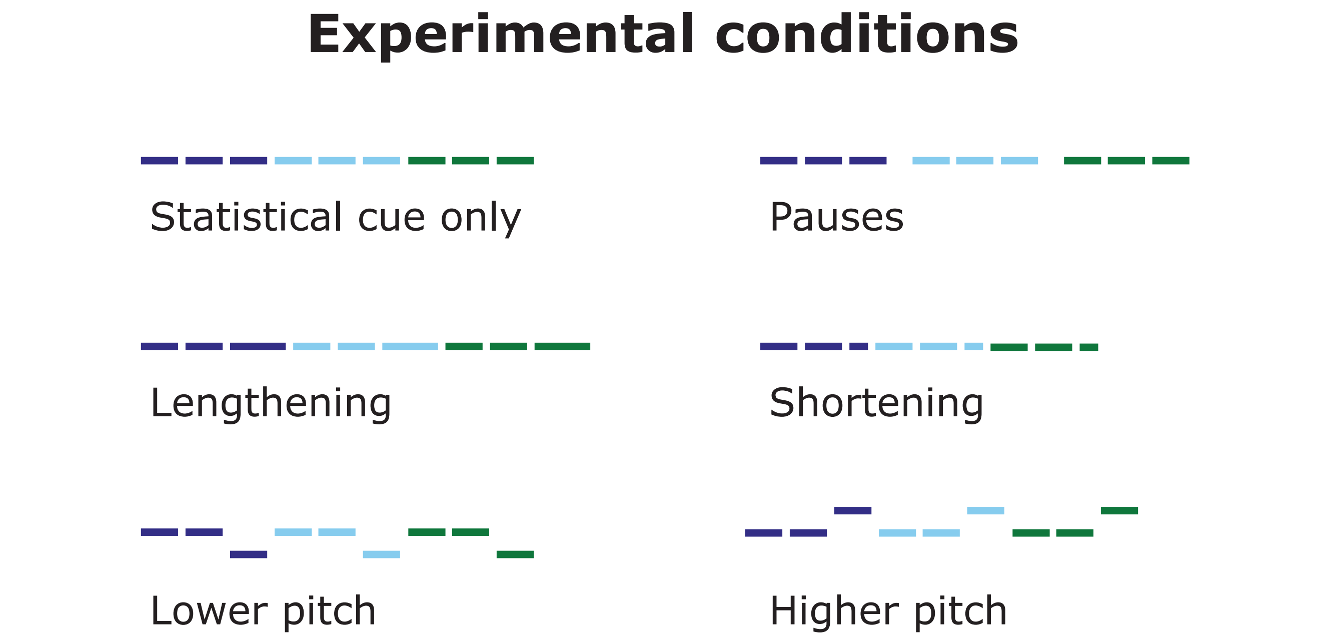


FIGURE 3 | Overview of the experimental conditions. For each condition, the figure shows an example speech-stream of three words. Lines denote syllables, line length indicates duration, and line height pitch. Colors denote statistical words.

4. Shortening condition. This condition was identical to the baseline condition, except that in addition to the statistical cues, the final syllable of each word was shortened by 50%, i.e., by the same proportion as syllables were lengthened in the lengthening condition, and thus had a duration of 250 ms.

5. Higher pitch condition. This condition was identical to the baseline condition, except that in addition to the statistical cues, the pitch of the final syllable of each word was increased to 260 Hz, making it 50 Hz higher than the pitch of all other syllables (cf. Thiessen and Saffran, 2003; Tyler and Cutler, 2009).

6. Lower pitch condition. This condition was identical to the baseline condition, except that in addition to the statistical cues, the pitch of the final syllable of each word was decreased to 160 Hz, making it 50 Hz lower than the pitch of all other syllables (as per the higher pitch condition).

EXPERIMENT 1: PILOT STUDY

Participants and Experimental Procedure

We tested 202 participants (19% male, mean age: 25.26), who were all native speakers of German and reported no auditory impairments. We used a between-subjects design: each participant was tested on one of six experimental conditions only (33 participants each in the pause, lengthening, and higher pitch condition; 34 participants each in the statistical cue only condition and the shortening condition; 35 participants in the lower pitch condition). Experimenters recruited the participants individually at the campus of the University of Vienna and they were tested *in situ* (e.g., in hallways, offices, public seating areas, etc.), while sitting or standing. Testing was performed with mobile testing equipment, i.e., a laptop computer and Sennheiser HD206 over-ear headphones. We ensured that the environment was free from obvious loud noise, but some background noise of other people walking by or chatting was unavoidable. We think that the effect of this background noise was minimal because participants could self-adjust the volume of the speech stream; none of them reported difficulties hearing the sounds.

Prior to the start of each experiment, participants were told that they would participate in an “Alien Language Learning Study” (as e.g., in Kirby et al., 2008), in which they would listen to a speech stream of an artificial pseudo-language and should decide for a set of 12 test stimuli whether they considered these to be “words” of the artificial language or not. Before listening to the speech stream, participants received a sheet of paper with all 12 test stimuli and were orally instructed in a standardized way to use a pen to circle the “words” of the “alien language” that they were about to hear. Participants listened to the speech stream for ∼1 min (see “Stimuli” below for the precise lengths) and rated the 12 test stimuli simultaneously. Typical exposure lengths in speech stream experiments conducted with adults in a laboratory are slightly longer, but because we tested in a natural environment, where it may be hard to concentrate during longer exposure times, we chose a shorter exposure time more typical for infant experiments (Saffran et al., 1996a; Thiessen and Saffran, 2003; Erickson and Thiessen, 2015) and compensated for this difficulty by using a rather low speech rate (see above; Song et al., 2010; Palmer and Mattys, 2016). These parameters were expected to facilitate speech segmentation in a natural environment. Including instructions, the overall experimental procedure lasted for ∼5 min. Immediately after participation, there was a short debriefing and participants’ questions about the study were answered. Participants received no financial reward.

TABLE 1 | Artificial words used for the different artificial pseudo-languages in the three experiments.

Experiment 1	Experiments 2 & 3			
Language 0	**Language 1**	**Language 2**	**Language 3**	**Language 4**
/batuki/	/bakupo/	/pifoke/	/dafego/	/mabopi/
/togabi/	/delaru/	/rovali/	/pebomi/	/veduka/
/punido/	/fumesi/	/nusema/	/kirune/	/sigale/
/dapiku/	/gonite/	/tabigu/	/lutiva/	/tonifu/

TABLE 2 | All possible part-words of pseudo-language 1, which consists of the words *bakupo*, *delaru*, *fumesi*, *gonite*.

Part-words 1–2				Part-words 2–1			
po-dela	ru-baku	si-baku	**te-baku**	kupo-de	**laru-ba**	mesi-ba	nite-ba
po-fume	**ru-fume**	si-dela	te-dela	**kupo-fu**	laru-fu	mesi-de	**nite-de**
po-goni	ru-goni	**si-goni**	te-fume	kupo-go	laru-go	**mesi-go**	nite-fu

Part-words that share the same word-initial syllables are grouped in columns. An example of one possible selection of part-words for the test phase is underlined.

Stimuli

The artificial pseudo-language consisted of four words with three CV (consonant-vowel) syllables each (**Table 1**, column 1, "Language 0"). The CV syllables were created from a pool of four vowels (a, u, i, o) and seven consonants (p, t, k, b, d, g, n). We ensured that the words created from this pool did not contain identical syllables, and were not existing words in German or English (which our participants spoke as a second language).

For the creation of the continuous speech streams of each condition, the four words were pseudo-randomly concatenated, with the restriction that no word could occur twice in a row. Each word was followed by each of the three remaining words equally often, which led to between-word transition probabilities of 0.33. One speech stream consisted of 40 words (i.e., each of the four words occurred 10 times in the stream). Depending on the condition, this led to total durations of the speech stream of 50 s (shortening condition), 60 s (baseline condition, lower pitch condition, and higher pitch condition), or 70 s (pause condition and lengthening condition).

The twelve test stimuli consisted of different *stimulus types*: four of the test stimuli were statistical *words*, i.e., the words that made up the particular artificial pseudo-language, and eight of the test stimuli were statistical *part-words*. Part-words could be of two different part-word classes and were created from syllables across word boundaries: either from the final syllable of a word and the initial and medial syllable of the following word (henceforth *part-words 1-2)*, or from the medial and final syllable of a word and the initial syllable of another word (henceforth *part-words 2-1*). Thus, crucially, in part-words 1–2, the original final syllables, which carried a prosodic cue in experimental conditions, occurred word-initially, and in part-words 2–1, the original final syllables occurred word-medially (see **Figure 1**). This procedure yielded 12 possible part-word stimuli in each part-word class (see **Table 2** e.g., of part-words of language 1). As actual test stimuli, we selected four different stimuli of each part-word class, namely /ku-toga/, /ki-puni/, /do-toga/, and /bi-dapi/ as part-words 1–2, and /tuki-pu/, /piku-ba/, /tuki-da/ and /nido-ba/ as part-words 2–1.

To create the actual sound stimulus, each syllable was recorded by a female native speaker of American English. Each syllable was recorded individually in order to avoid co-articulation between syllables within a word (coarticulation could serve as an additional cue to speech segmentation, as e.g., in Johnson and Jusczyk, 2001, modifying the effects of the individual prosodic cues). The acoustic parameters of each syllable were modified using Praat (version 6.0.36; Boersma and Weenik, 2017), and the output syllables were then concatenated using custom code written in Python 3.6.3 to create the speech streams.

The acoustic modifications of the syllables concerned their fundamental frequency ("pitch"), duration, and amplitude. Pitch and duration of the syllables were modified using the pitch-synchronous overlap add (PSOLA) algorithm, which is a signal processing technique used for speech processing and synthesis implemented in Praat (Moulines and Charpentier, 1990). We used customized Praat scripts, which were based on the Praat functions "Manipulate→Replace Pitch Tier" and "Manipulate→Replace Duration Tier" to change syllable pitch and duration. For each syllable in the baseline condition, the fundamental frequency (f_0) was normalized to a mean of 210 Hz, and the duration of each syllable was normalized to a mean of 500 ms. Durational changes were applied to the entire syllable except for the first 20 ms. This was done to avoid changes in voice onset time and associated consonant shifts. For the experimental conditions, all syllables were manipulated according to the same procedure to meet the respective duration and pitch specifications (see chapter 2.2). Syllable amplitude was made consistent by scaling the amplitude of each syllable so that its absolute peak amplitude was 0.99 (in Praat: Sound→ Modify→Scale peak→New absolute peak: 0.99).

To avoid possible cueing to word boundaries, the continuous speech streams had a gradual fade-in and fade-out over the first and last five words, respectively, so that the perceived start and the end of the speech stream did not align with word boundaries. For the fade-in, the amplitude of the first 15 syllables, i.e., of each syllable of the first five words of the stream, was increased by 6.66% of the peak amplitude, so that at the beginning of the sixth word, the full amplitude was reached. Similarly, for the fade-out, we decreased the amplitude of each of the last 15 syllables by 6.66% of the peak amplitude. Amplitude manipulation was implemented in Python and Praat.

D' Analysis and Results

To obtain a general overview of the influence of experimental conditions on participants' discrimination of words and non-words, we used signal detection theory measures and calculated d' values (Green and Swets, 1966; Macmillan and Kaplan, 1985; Macmillan and Creelman, 2005), based on hit rates (i.e., selection of statistical words as words) and false alarm rates (i.e., selection of statistical part-words as words). Perfect performance (100% hits and 0% false alarms) causes mathematical problems in signal detection theory, requiring *post-hoc* changes to these values to avoid divide-by-zero issues when calculating d prime values. Therefore, we adjusted perfect hit rates and false alarm rates according to the standard 1/(2N) rule, which adds 1/(2N) to proportions of 0 and subtracts 1/(2N) from proportions of 1 (Hautus, 1995; Stanislaw and Todorov, 1999; Brown and White, 2005; Macmillan and Creelman, 2005). D' values of 0 indicate that participants selected words and non-words at chance level, d' values above 0 indicate a discrimination performance above chance (i.e., participants perceived many statistical words as words), and d' values below 0 indicate a discrimination performance worse than chance (i.e., participants perceived many statistical part-words as words). We computed 95% confidence intervals (CIs; **Figure 4**) to determine if the differences between individual groups and the differences to chance level performance were significant. Confidence intervals that do not overlap with each other indicate significant differences between groups. Confidence intervals that do not include d' values of 0 indicate that word perception is either better (CIs above 0) or worse (CIs below 0) than chance (Cumming and Finch, 2005; Cumming, 2012, 2014).

Our calculation of d' values revealed that in the baseline condition, participants performed better than chance (**Figure 4**), indicating that statistical cues alone sufficed to detect words in the speech stream. In the pause and lengthening conditions, participants excelled on the task, indicating that pauses and final lengthening provided strong convergent cues for speech segmentation. In the pause and lengthening conditions, the participants' performance was also significantly higher than in the baseline condition, indicating that adding these cues to a speech stream significantly facilitates segmentation performance.

In contrast, in the lower and higher pitch conditions, participants showed only moderate discrimination performance, which was above chance but did not significantly differ from the baseline statistical condition. This suggests that enhancing statistical cues with a pitch modification on the word-final syllable did not appreciably aid speech segmentation for our listeners. Interestingly, in the shortening condition, the performance was in fact very poor and significantly worse than baseline, showing that shortening final syllables hindered word segmentation. This suggests that prosodic cues can override statistical cues when they conflict, but does not yet show if the low performance was due to participants perceiving the shortened syllable in word-initial (part-words 1–2) or word-medial (part-words 2–1) position. To clarify this, we conducted a more fine-grained analysis involving a generalized linear mixed model.

Generalized Linear Mixed Model: Analysis

To investigate if the different prosodic cues had an effect on which stimulus type (statistical word or one of the two statistical part-word types) the participants perceived, i.e., on the "correctness" of their responses on the three different stimulus types, we fitted a logistic Generalized Linear Mixed Model (Baayen, 2008) with logit link function (McCullagh and Nelder, 1989). *Condition* and *stimulus type*, as well as their interaction, were included as fixed effects into the model. We also entered a random intercepts effect of *participant* in the model. To avoid inflated type I error rates we included a random slope (Schielzeth and Forstmeier, 2009; Barr et al., 2013) of *stimulus type* within *participant*. Before including this factor into the random slope we manually dummy coded and then centered it. The sample size for this model was 2,424 data points (202 individuals tested on one condition each, with 12 trials), 1,719 of which were correct responses. Responses were coded as "correct" when participants selected the statistical words as being "words" and rejected the statistical part-words as being "words" of the artificial language, so that for each stimulus type, perfect performance would be 100%, and chance-level performance (guessing) would be 50% correct responses.

The model was fitted in R (version 3.6.0; R Development Core Team, 2018), using the function *glmer* of the R-package *lme4* (version 1.1.21; Bates et al., 2015) and the optimizer "bobyqa".

To test the overall significance of *condition* (i.e., its main effect and its potential interaction with *stimulus type*), we used a likelihood ratio test to compare our full model to a null model that was identical to the respective full model except for that it did not include *condition* and its interaction with *stimulus type* (R function *anova* with argument "test" set to "Chisq"; Dobson, 2002).

P-values for the effect of individual predictors are based on likelihood ratio tests that compare the full model with respective reduced models lacking the effects one at a time (R function *drop1*; Barr et al., 2013). We determined model stability by dropping individuals one at a time and comparing the estimates obtained for these subsets with those obtained for the full data set, which revealed that our model was fairly stable (see **Supplementary Table 1**). We determined confidence intervals of estimates and the fitted model using a parametric bootstrap (function *bootMer* of the package *lme4*, using 1,000 parametric bootstraps).

Generalized Linear Mixed Model: Results

Overall, the full model (for details, see **Supplementary Table 1**) was significantly different from the null model, indicating an effect of *condition* or its potential interaction with *stimulus type* on the perception of words in a speech stream (likelihood ratio test: $\chi^2 = 147.865$, df = 15, $p < 0.001$). Word perception was measured by the proportion of "correct" answers in the experiment, specifically, the proportion of statistical words and part-words that listeners identified as words and part-words, respectively. More specifically, we found that the interaction between *condition* and *stimulus type* had a significant effect on word perception (likelihood ratio test: $\chi^2 = 63.129$, df = 10, $p < 0.001$), indicating that the pattern of correct responses to words vs. part-words varied between conditions (see **Figure 5**). The computed confidence intervals (**Figure 5**) allow us to make comparisons between individual groups. This confirms the main results from the d' analysis above, and additionally allows

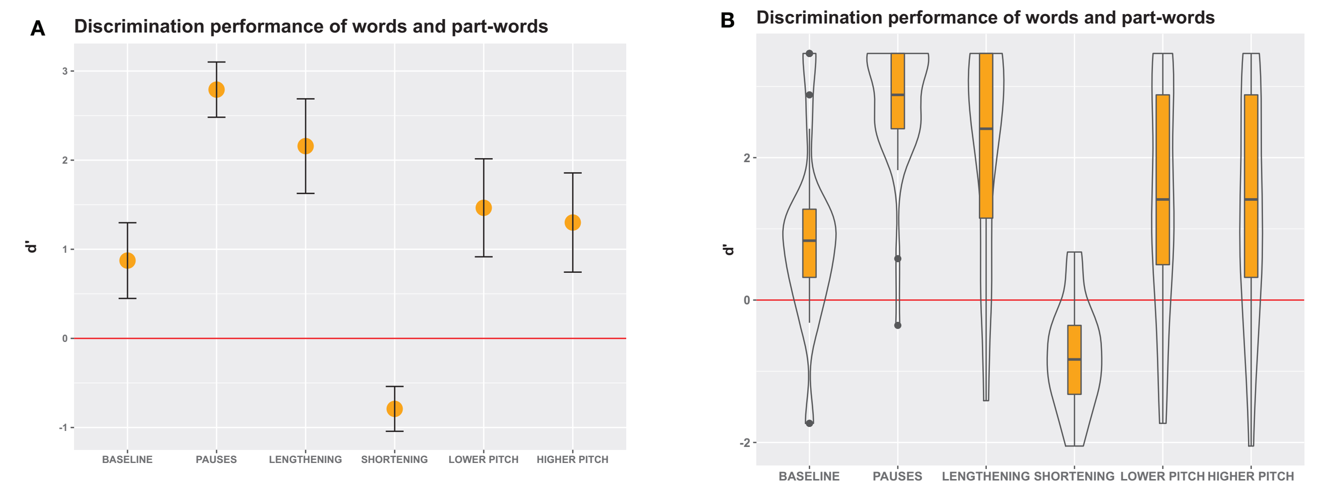


FIGURE 4 | D' measures in Experiment 1, an immediate-decision task where participants decided whether visually presented letter strings were words that they had perceived in the speech stream or not. **(A)** Mean and 95% confidence intervals of participants' responses. Non-overlapping confidence intervals indicate significant differences between the groups. Confidence intervals that do not include 0 indicate significant differences from chance performance. **(B)** Boxes depict medians and quartiles, whiskers minimum and maximum values, and black dots outliers. Violin shapes around the boxes depict the distribution of d' values. The width of the violin shapes at a given y coordinate corresponds to the number of d' values in this region. Red lines: chance level performance.

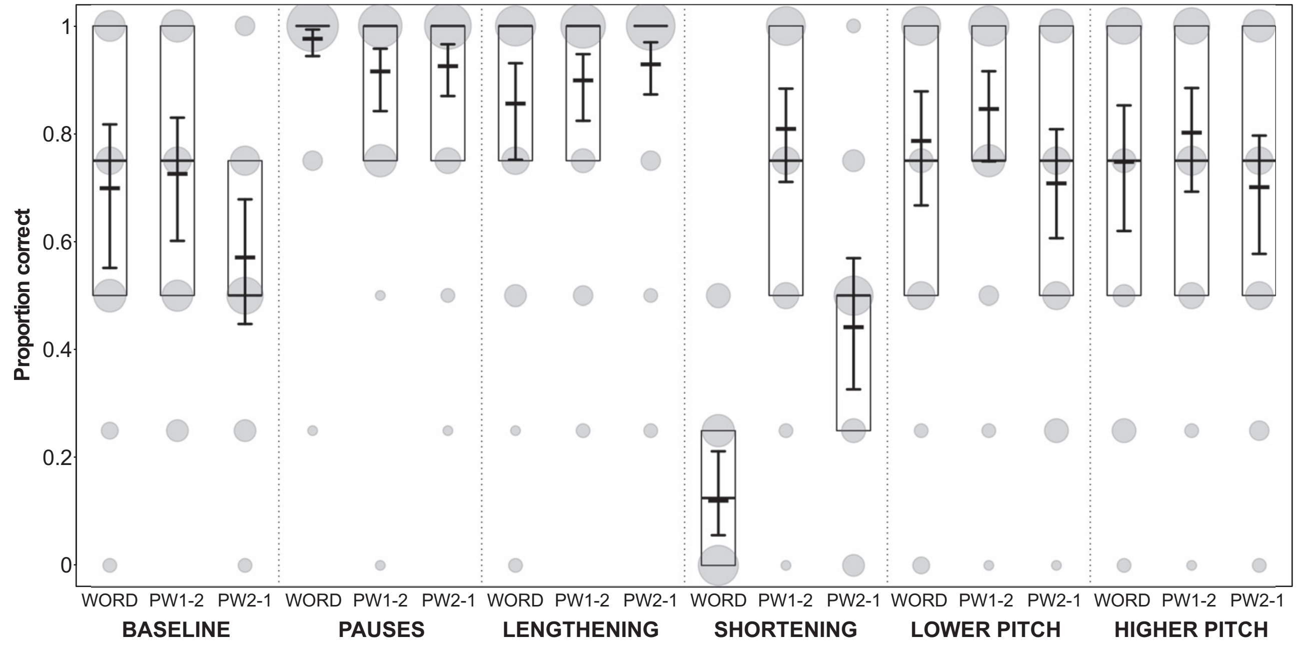


FIGURE 5 | Proportion of participants' correct answers in Experiment 1. Proportions are displayed for each condition and each stimulus type (WORD, word with modified syllable in final position; PW1-2, part-word with modified syllable in initial position; PW2-1, part-word with modified syllable in medial position). Model results: thick horizontal black lines, with error bars depicting the bootstrapped 95% confidence intervals. Boxes depict medians and quartiles, and gray dots the actual observations (the area of the dots indicates the number of responses per combination of condition, stimulus type, and proportion correct).

comparisons between participant performance on the three different stimulus types.

In all conditions except the shortening condition, the performance on words, part-words 1–2 and part-words 2–1 was very similar (**Figure 5**), i.e., words were correctly selected as being statistical words and part-words were correctly rejected. Interestingly, in the shortening condition, our analysis (see model estimates and confidence intervals in **Figure 5**) revealed very clearly that performance on part-words 2–1 (i.e., stimuli where the shortened syllables occurred word-medially) was poor, because participants identified many part-words 2–1 as *words* (which count as false alarms in our analysis). That is, they interpreted the shortened syllables as being word-medial. This is a violation of typical German word stress because German word-stress usually occurs word-medially, and stressed syllables are usually lengthened (see **Figure 1**). Furthermore, performance on the statistically correct words (where the shortened syllables occurred word-finally) was also very poor, because participants incorrectly identified many of these words as *part-words*. However, they correctly identified most part-words 1–2 (i.e., stimuli where the shortened syllables occurred word-initially) as *part-words*. This clearly shows that in this condition participants were biased to perceive as "words" those stimuli where the duration of the medial syllable was shortened.

EXPERIMENT 2

The main aim of this experiment was to replicate Experiment 1 in a more controlled laboratory setting. For the sake of this comparison, we kept the key aspects of Experiment 1, most notably that participants evaluated the test items on a sheet of paper while listening to the speech stream, but Experiment 2 was a within-subjects study that controlled more aspects of the experimental procedure via randomization than Experiment 1. Experiment 2 specifically focused on the conditions that significantly differed from the baseline in Experiment 1, namely conditions 1 to 4, and omitted the pitch manipulation.

Participants and Experimental Procedure

We tested 34 participants (21% male, mean age: 24.85), who were all native speakers of German and reported no auditory impairments. Participants were recruited via posters or online advertisements. Participant instructions and the overall testing procedure were identical to Experiment 1, except that participants were now individually tested in a quiet laboratory setting. While sitting ∼60 cm from a 13$''$ monitor, they were shown instructions and listened to the speech stream via an experimental interface created in PsychoPy (version 1.90.3; Peirce, 2007). Further, we used a within-subjects design, in which all participants were tested on all four conditions in a randomized order. The speech stream of each condition now lasted twice as long, for ∼2 min (see "Stimuli" below for details). Between each condition, participants were given a 30 s break. No feedback on the responses was provided. Thus, including instructions and a final debriefing, the experiment lasted ∼20–25 min. Participants were given modest monetary compensation for their participation.

Stimuli

Because each participant was tested on four different experimental conditions, we created four different artificial pseudo-languages (**Table 1**, columns 2–5), consisting of four words with three CV (consonant-vowel) syllables each. For each participant, we pseudo-randomized which pseudo-language was used for which condition. We carefully controlled stimulus creation to avoid potential transfer or priming from words learned in one condition in one pseudo-language to words in another condition in another pseudo-language. Therefore, the CV syllables were created from a pool of five vowels (a, e, i, o, u) and 13 consonants, namely six stops (b, d, g, p, t, k), three fricatives (f, v, s), and four sonorants (m, n, l, r). In total, the four words of each language required 12 vowels and 12 consonants. To minimize possible cues resulting from the distribution of vowels and consonants we ensured that within each pseudo-language used in Experiments 2 and 3, vowels were evenly distributed (two of the vowels occurred three times and three of the vowels occurred twice) and that no word contained the same vowel twice. Also, no consonant occurred within one pseudo-language more than once. Thus, each syllable was unique within a pseudo-language. Moreover, across all four pseudo-languages, none of the syllables occurred more than twice, with the majority of the syllables only occurring once.

One speech stream consisted of 96 words (i.e., each of the four words occurred 24 times in the stream). Depending on the condition, this led to total durations of the speech stream of 120 s (shortening condition), 144 s (baseline condition), or 168 s (pause condition and lengthening condition).

As in Experiment 1, each participant received 12 test stimuli per condition, which consisted of statistical *words* and statistical *part-words*, created as described above for Experiment 1. For each participant and in each condition, the set of test stimuli included four statistical *words*. The four *part-words 1–2* and the four *part-words 2–1* were pseudo-randomly selected for each individual participant and each condition. We ensured that each first and second part was represented once in each part-word class (e.g., see words highlighted in bold in **Table 2**).

The actual sound signals of the speech streams were created as in Experiment 1, except that in this experiment the syllables from which the speech streams were created were recorded by a different female native speaker (in this case of German).

D' Analysis and Results

Our calculation of d' values (for details about the analysis, see Experiment 1) revealed that discrimination performance was best in the pause condition, moderately good in the lengthening condition and almost above chance in the baseline condition. Shortening again hindered speech segmentation compared to the baseline (**Figure 6**). Thus, the effects were similar to those in Experiment 1, but performance was worse. As in Experiment 1, we performed a generalized linear mixed model to investigate the reasons for the low performance in the shortening condition.

Generalized Linear Mixed Model: Analysis

As in Experiment 1, we fitted a logistic Generalized Linear Mixed Model (Baayen, 2008) with logit link function (McCullagh and Nelder, 1989) to test whether the perception of words in the speech stream was influenced by condition and stimulus type (statistical word or one of the two statistical part-word types). We again included *condition* and *stimulus type*, as well as their interaction as fixed effects into the model. To control for the effects of *pseudo-language* (factor with four levels; participants were exposed to a different pseudo-language in each of the four

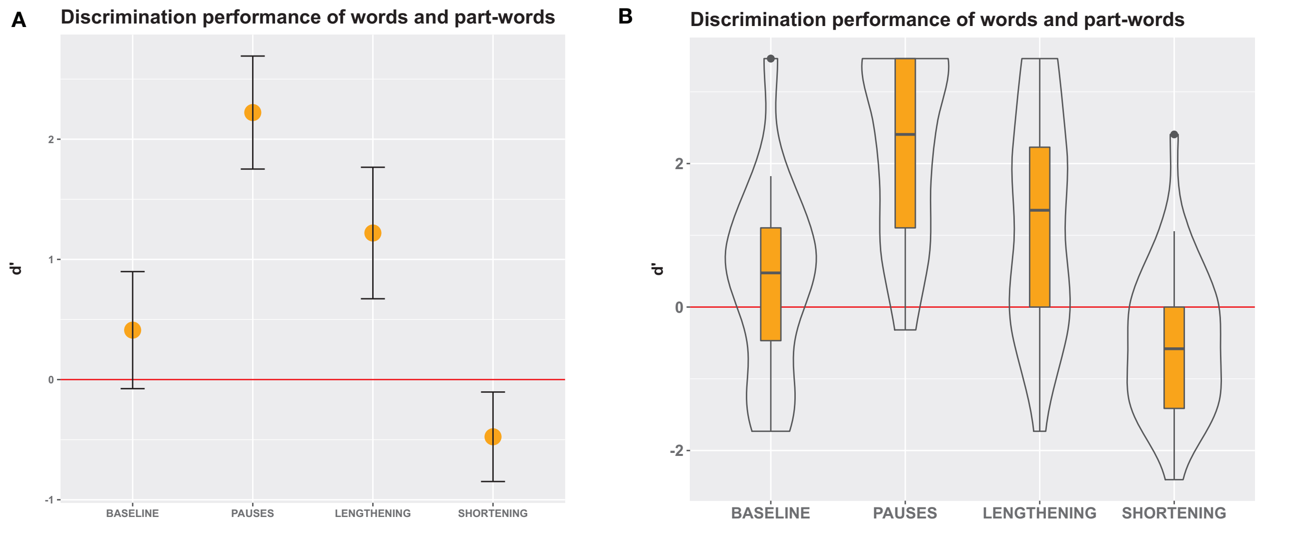


FIGURE 6 | D' measures in Experiment 2, an immediate-decision task where participants decided whether visually presented letter strings were words that they had perceived in the speech stream or not. **(A)** Mean and 95% confidence intervals of participants' responses. Non-overlapping confidence intervals indicate significant differences between the groups. Confidence intervals that do not include 0 indicate significant differences from chance performance. **(B)** Boxes depict medians and quartiles, whiskers minimum and maximum values, and black dots outliers. Violin shapes around the boxes depict the distribution of d' values. The width of the violin shapes at a given y coordinate corresponds to the number of datapoints in this region. Red lines: chance level performance.

conditions) and *order of the conditions* (covariate with values 0–3), we included them as further fixed effects. We also entered a random intercepts effect of *participant* in the model. Again, to keep type I error rates at the nominal level of 0.05, we included random slopes (Schielzeth and Forstmeier, 2009; Barr et al., 2013) of *condition*, *stimulus type*, their interaction, *order of the conditions,* and *language* within *participant*. Before including factors into the random slopes we manually dummy coded and then centered them. We did not include the correlations between random intercept and random slopes terms in the final model because an initial model including these correlations and thus being maximal with regard to random effects failed to converge. The control predictor *order of the conditions* was z-transformed (to a mean of zero and a standard deviation of one). The sample for this model was 1,632 data points (34 individuals tested on four conditions with 12 trials each), 1,066 of which were correct responses.

Significances of the individual predictors, model stability (for details see **Supplementary Table 2**) and confidence intervals were calculated as described for Experiment 1.

Generalized Linear Mixed Model: Results

In experiment 2, a comparison of the full model with the null model again revealed an effect of either *condition* or its potential interaction with *stimulus type* on the perception of words in a speech stream (likelihood ratio test comparing the full and the null model: $\chi^2 = 63.00$, df $= 9$, $p < 0.001$; for model details, see **Supplementary Table 2**). Exploring these effects, we found that the interaction effect between *condition* and *stimulus type* was non-significant (likelihood ratio test: $\chi^2 = 11.329$, df $= 6$, $p = 0.079$). However, because this interaction effect was very close to being significant, it is not justified to exclude it from the model and determine the effect of *condition* alone. Overall, these results again reflect different response patterns between conditions (see **Figures 6**, **7**), but the differences between the conditions were not as prominent as in Experiment 1. Again, this confirms the main results from the d' analysis above. Although the interaction effect did not meet the threshold for statistical significance, comparisons between the three different stimulus types can shed light on the speech segmentation strategies employed in the different conditions and provide valuable comparison to experiments 1 and 3. With regard to the outcomes of experiment 1, we were most interested in the shortening condition, for which we predicted a low performance on words and part-words 2–1, and a high performance on part-words 1–2.

The comparison between the performances on the three different stimulus types (see model estimates and confidence intervals in **Figure 7**) revealed that in the pause condition, participants showed high performance on all stimuli (correctly identifying statistical words as *words*, and statistical part-words as *part-words*). In the baseline and the lengthening condition, participants performed rather well at identifying part-words, but relatively poorly identifying words. This indicates a bias to select only a few stimuli as words, leading to a considerable number of misses for words. In the shortening condition, participants again performed worst, missing many words, and labeling them as part-words incorrectly (see model estimates and confidence intervals in **Figure 7**), indicating cue conflict for this condition. The performance on part-words 1–2 and part-words 2–1 was similar, which indicates that participants perceived the

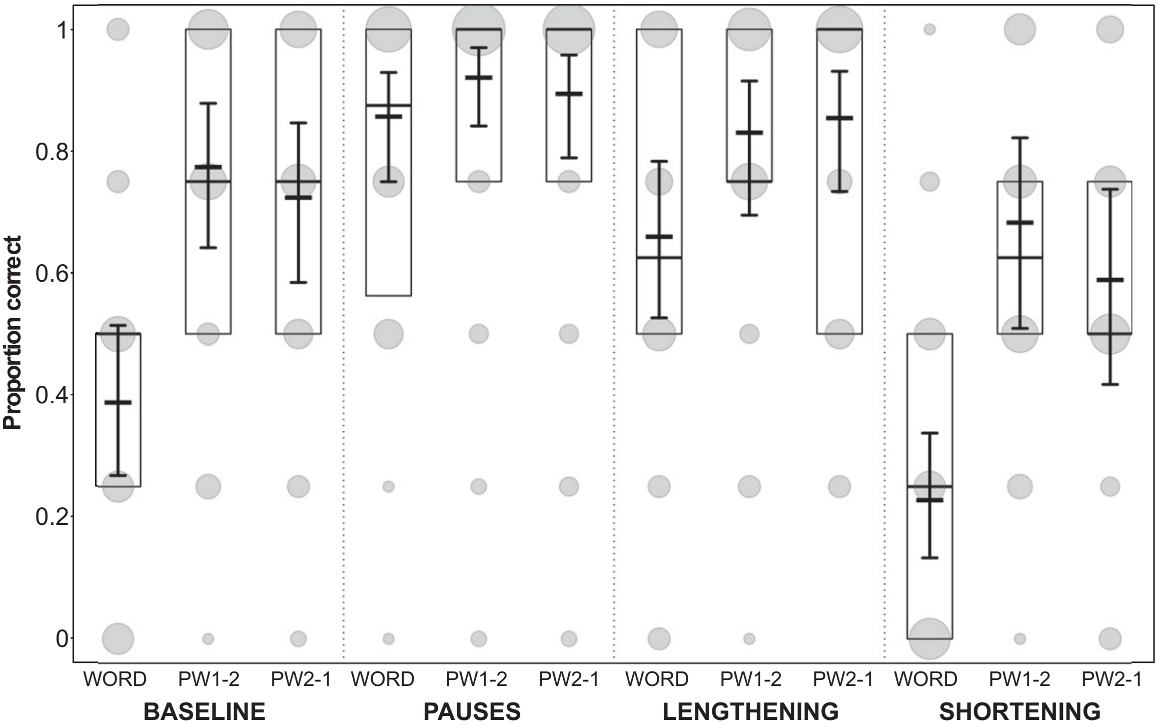


FIGURE 7 | Proportion of participants' correct answers in Experiment 2. Proportions are displayed for each condition and each stimulus type (WORD, word with modified syllable in final position; PW1-2, part-word with modified syllable in initial position; PW2-1, part-word with modified syllable in medial position). Model results: thick horizontal black lines, with error bars depicting the bootstrapped 95% confidence intervals. Boxes depict medians and quartiles, and gray dots the actual observations (the area of the dots indicates the number of responses per combination of condition, stimulus type, and proportion correct).

shortened cue on the word-medial and word-initial syllable equally often.

The control predictors *order of the conditions* (likelihood ratio test: $\chi^2 = 0.945$, df $= 1$, $p = 0.329$) and *pseudo-language* (likelihood ratio test: $\chi^2 = 1.725$, df $= 3$, $p = 0.631$) did not have a significant effect on discrimination performance of words and part-words. The null effect of the predictor *order of the conditions* indicates that there was no cross-condition interference of segmentation strategies and participants did not infer a consistent rule that they transferred from condition to condition.

EXPERIMENT 3

Given the overall consistent results of Experiments 1 and 2, the main goal of Experiment 3 was to probe their robustness, by modifying the paradigm. In particular, we added a more pronounced auditory memory component by delaying responses and presenting the test stimuli acoustically instead of visually. Participants first listened to the entire speech stream. Then, in a subsequent test phase, they listened to single probe stimuli and made a decision for each stimulus whether it was a word or a part-word. Correct responses thus required participants to remember any words that they perceived during presentation, despite interference from repeatedly hearing part-words during testing. Thus, Experiment 3 tested not just the effect of our manipulations on the immediate perception of test stimuli, but also how well people remembered them. This makes this experiment resemble real-life language learning more closely, and resembles many previous speech segmentation experiments (e.g., Tyler and Cutler, 2009; Frost et al., 2017; Ordin et al., 2017). In Experiment 3, we investigated all six experimental conditions from Experiment 1.

Participants and Experimental Procedure

We tested 42 participants (26% male, mean age: 24.19 years), who were all native speakers of German and reported no auditory impairments. Participants were recruited via posters or online advertisements. As in Experiment 2, testing happened in a laboratory; the experiment was administered via an experimental interface created in PsychoPy (version 1.90.3; Peirce, 2007), which coordinated the presentation of instructions, speech streams and acoustic test stimuli, and collected key-press responses. We used a within-subjects design, in which all participants were tested on four of the six experimental conditions, namely the baseline condition, the pause condition, one of the durational cue conditions (either the lengthening or the shortening condition) and one of the pitch cue conditions (either the lower or higher pitch condition). Which of the durational and pitch cue conditions a participant ran was pseudo-randomized. We did not test participants on all six conditions to reduce the chance that they inferred a rule (e.g., "the modified syllable is always the last syllable of the word") that might transfer from condition to condition. The presentation order of the conditions was randomized. Immediately after listening to each speech stream, participants listened to the corresponding 12 test stimuli in a randomized order and indicated, after each stimulus, whether they considered it to be a word in the preceding artificial language or not. Participants pressed a green-labeled key on a computer keyboard to indicate "word" and a red key if not. One half of the participants pressed the green key with the left hand and the red key with the right hand. To avoid effects of handedness, for the other half of the participants, this was reversed. No feedback on the responses was provided.

As in Experiment 2, the speech stream for each condition lasted for ∼2 min (see "Stimuli" below for details), participants completed each test phase at their own pace, and between the conditions, participants were given a 30 s break. Thus, including instructions and a final debriefing, the experiment lasted ∼20–25 min. Participants were given modest monetary compensation for their participation in the experiment.

Stimuli

For Experiment 3, we used the same artificial languages (**Table 1**, columns 2–5), the same speech streams (including two additional speech streams for the two pitch conditions) and the same test stimuli as for Experiment 2 (e.g., see words highlighted in bold in **Table 2**). The acoustic versions of the test stimuli were created from syllables spoken by the same female native speaker of German as Experiment 2, in the same way as previous speech streams (see "Stimuli" in Experiment 1 and 2). All syllables were normalized to the default length of 500 ms and the default pitch of 210 Hz. The test stimuli did not carry any modifications of duration or pitch from these standards.

D' Analysis and Results

Our calculation of d' values (for details about the analysis, see Experiment 1) revealed that pauses again significantly improved discrimination performance compared to the baseline (**Figure 8**). In all other conditions, discrimination performance was near chance level, except for the higher pitch condition, where it was slightly above chance. There was a tendency that participants discriminated words and part-words better than chance in the baseline and lengthening conditions and worse than chance in the shortening and lower pitch conditions. Thus, the directions of the effects were similar to Experiments 1 and 2, but all effects besides those of pauses were very weak. As in Experiments 1 and 2, we performed a generalized linear mixed model to investigate the reasons for the generally low performance.

Generalized Linear Mixed Model: Analysis

As for Experiments 1 and 2, we fitted a logistic Generalized Linear Mixed Model (Baayen, 2008) with logit link function (McCullagh and Nelder, 1989) to test whether the perception of words in the speech stream was influenced by condition and stimulus type (statistical word or one of the two statistical part-word types). *Condition* and *stimulus type*, as well as their interaction were included as fixed effects into the model. To control for the effects of *pseudo-language* (factor with four levels; participants were exposed to a different pseudo-language in each of the four conditions), *order of the conditions* (covariate with values 0–3), and *trial number* (counting from 0 to 11 within each condition), these were included as additional fixed effects. The predictor *pseudo-language* was manually dummy coded with Language 1 being the reference category, and then centered. As in previous experiments, we entered a random intercept of *participant* in the model, and included random slopes (Schielzeth and Forstmeier, 2009; Barr et al., 2013) of *condition*, *stimulus type*, their interaction, *order of the conditions* and *trial number* within *participant*. Again, before including factors into the random slopes we manually dummy coded and then centered them. The correlations between random intercept and random slopes terms were not included in the final model because an initial model including these correlations—and thus being maximal with regard to random effects—did not converge. The control predictors *order of the conditions* and *trial number* were z-transformed (to a mean of zero and a standard deviation of one). The sample for this model consisted of 42 individuals tested on 4 conditions with 12 trials each. This yielded 2016 data points in total, 1,063 of which revealed a correct response.

Significances of the individual predictors, model stability (for details see **Supplementary Table 3**) and confidence intervals were calculated as described for Experiment 1.

Generalized Linear Mixed Model: Results

As for Experiments 1 and 2, the comparison of the full model (for details, see **Supplementary Table 3**) and the null model for Experiment 3 revealed that *condition* or its potential interaction with *stimulus type* had an impact on the perception of words (likelihood ratio test: $\chi^2 = 62.20$, df $= 15$, $p < 0.001$). Unpacking these effects, we found that the interaction between *condition* and *stimulus type* had a significant effect on word perception (likelihood ratio test: $\chi^2 = 31.963$, df $= 10$, $p < 0.001$). This means that the pattern of correct responses on words and part-words varied between conditions (see **Figures 8**, **9**). However, the overall results of Experiment 3 were slightly less clear than for Experiments 1 and 2. This confirms the main results from the d' analysis above.

Participants showed quite high performance on words and both part-word types in the pause condition, but they showed a high performance on words and a low performance on part-words in the baseline, lengthening and high pitch conditions. The rather low performance on part-words in these conditions indicates that participants had the tendency to select many incorrect stimuli as words, resulting in many false alarms for part-words. In the shortening and low pitch conditions, performance was rather low on words and part-words. Interestingly, as in Experiment 1, in the shortening condition, performance on part-words 2–1 was low, which indicates that participants had the tendency to perceive stimuli where shortening happened on the medial syllable as words (see model estimates and confidence intervals in **Figure 9**).

The control predictors *pseudo-language* (likelihood ratio test: $\chi^2 = 4.013$, df $= 3$, $p = 0.260$), *order of the conditions* (likelihood ratio test: $\chi^2 = 0.159$, df $= 1$, $p = 0.692$), and *trial number* (likelihood ratio test: $\chi^2 = 0.014$, df $= 1$, $p = 0.907$) did

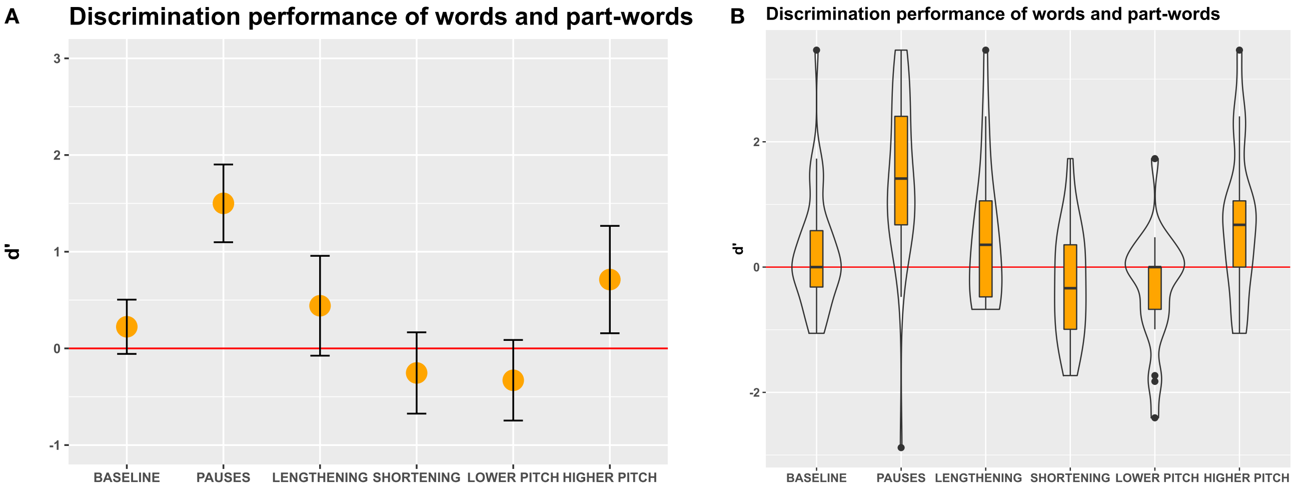


FIGURE 8 | D' measures in Experiment 3, a decision task where participants decided whether acoustically presented stimuli were words that they had perceived in the speech stream or not. **(A)** Mean and 95% confidence intervals of participants' responses. Non-overlapping confidence intervals indicate significant differences between the groups. Confidence intervals that do not include 0 indicate significant differences from chance performance. **(B)** Boxes depict medians and quartiles, whiskers minimum and maximum values, and black dots outliers. Violin shapes around the boxes depict the distribution of d' values. The width of the violin shapes at a given y coordinate corresponds to the number of d' values in this region. Red lines: chance level performance.

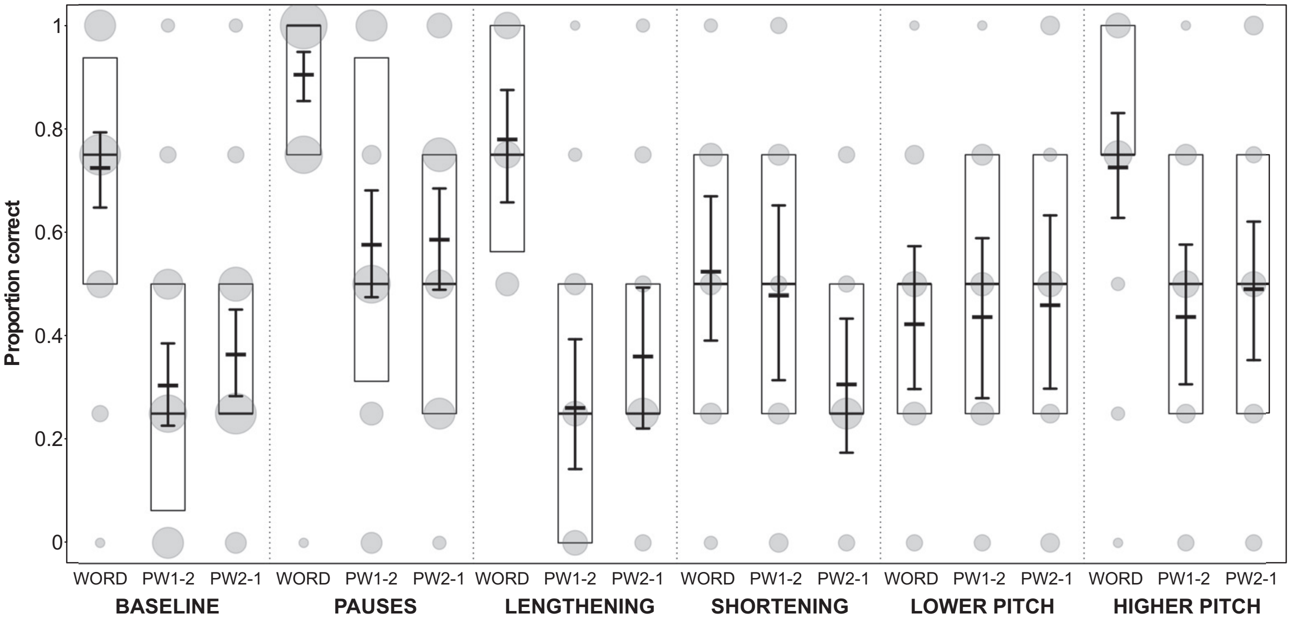


FIGURE 9 | Proportion of participants' correct answers in Experiment 3. Proportions are displayed for each condition and each stimulus type (WORD, word with modified syllable in final position; PW1-2, part-word with modified syllable in initial position; PW2-1, part-word with modified syllable in medial position). Model results: thick horizontal black lines, with error bars depicting the bootstrapped 95% confidence intervals. Boxes depict medians and quartiles, and gray dots the actual observations (the area of the dots indicates the number of responses per combination of condition, stimulus type, and proportion correct).

not have a significant effect on discrimination performance of words and part-words. The null effect of the predictor *order of the conditions* indicates that there was no cross-condition interference of segmentation strategies and participants did not infer a consistent rule that they transferred from condition to condition.

DISCUSSION

Our study indicates that manipulating prosodic information has clear effects on speech segmentation by adult German-speaking listeners, mostly improving performance relative to a statistics-only baseline (see non-overlapping confidence intervals in **Figures 4**, **6**, **8**). This basic result is consistent with considerable previously published data. A significant interaction between *condition* and *stimulus type* in Experiments 1 and 3 (and near significant interaction with $p = 0.079$ in Experiment 2) clearly shows that listeners' identifications of words and part-words differed in the different prosodic modification conditions. Our prosodic modifications occurred either on the final syllable of a trisyllabic nonsense word (for durational and pitch cues), or after it (for pauses). Our results further show that listeners interpreted different prosodic modifications as occurring at different positions in these trisyllabic words. This provides clear evidence that different prosodic cues have differing effects on speech segmentation, in an experiment where for the first time multiple prosodic cues were contrastively manipulated with other acoustic factors being closely controlled.

The Positive Effects of Pauses and Final Lengthening on Speech Segmentation

Overall, adding pauses and lengthening the final syllable converged with the statistical cues, significantly facilitating speech segmentation based on statistical cues alone. In Experiment 1, participants identified most of the test items correctly in the pause and lengthening condition, whereas in the baseline condition with statistical cues alone, performance was only slightly above chance (**Figure 4**). In Experiment 2, pause and lengthening cues improved identification of words, but not the rejection of part-words, compared to the baseline condition (see non-overlapping CIs in **Figure 7**). In Experiment 3, which added a pronounced memory component, pauses, but not final lengthening, led to a higher performance compared to the statistical-cues-only condition (**Figure 8**). This overall convergent effect of final lengthening is consistent with the language-universal occurrence of domain-final lengthening, but not a language-specific stress pattern because German trisyllabic words typically do not carry stress on their final syllables (for a discussion, see Crowhurst, 2016). Overall, these results are in accordance with a large body of previous research showing that final lengthening cues are perceived as converging with statistical cues, thus facilitating speech segmentation (Saffran et al., 1996b; Tyler and Cutler, 2009; Ordin and Nespor, 2016; Frost et al., 2017; Ordin et al., 2017). These results are thus consistent with the cue convergence hypothesis for pause and final lengthening cues.

The Negative Effect of Final Shortening on Speech Segmentation

In contrast, shortening the final syllable actively hindered the identification of statistical words, compared to statistical cues alone, consistent with the cue conflict hypothesis for final shortening cues. This was illustrated most clearly in Experiments 1 and 2, where identification of statistical words in the shortening condition was significantly lower than in the baseline condition (see **Figures 4**, **6**). In Experiment 3, overall performance in the shortening condition was quite low because either "correct" statistical words were missed, or part-words were mistakenly selected as "words" (see **Figures 8**, **9**). Interestingly, in Experiments 1 and 3, participants selected many part-words 2–1 as words (see low performance on part-words 2–1 in Experiments 1 and 3; **Figures 5**, **9**), indicating that participants tended to perceive shortened syllables as occurring word-medially. Thus, when prosodic and statistical cues conflict, prosodic cues overrode statistical cues in the speech segmentation process, yielding "word" percepts based on prosodic patterns that conflict with those based on transition probabilities. Prosody has also overpowered statistics in previous studies: English infants grouped syllables with a combination of longer duration, higher pitch and higher intensity as word-initial, disregarding statistical cues (e.g., Johnson and Jusczyk, 2001; Thiessen and Saffran, 2003; Johnson and Seidl, 2009). In our study, however, neither final lengthening nor pitch increase overrode statistical cues when occurring individually (although final lengthening significantly augmented such cues), but shortening alone sufficed to override the statistical cues.

Final shortening may strongly influence speech segmentation because listeners have a language-universal preference for final lengthening (Tyler and Cutler, 2009; Fletcher, 2010; but also: Ordin et al., 2017; White et al., 2020). Encountering the opposite cue—final shortening—might thus actively interfere with word segmentation. The contrasting results we observed for final lengthening and shortening cues were consistent with our hypothesis that contrasting cues should have contrasting effects. Participants may also have perceived the shortening cues on the medial syllables because, when medial syllables are short, final syllables are perceived as longer, which would again fit the language-universal preference for final lengthening. Another potential explanation for the word-medial perception of shortened syllables might be that some German trisyllabic words do carry stress on the initial or final syllable (Domahs et al., 2014; Ordin and Nespor, 2016; Ordin et al., 2017) and that in these words, medial syllables may appear weaker and shortened. However, overall, it seems unlikely that language-specific word stress patterns explain why shortening was perceived on medial syllables because German trisyllabic words are typically stressed on the medial syllable (Domahs et al., 2014; Ordin and Nespor, 2016; Ordin et al., 2017), and shortening is not typically associated with stress (Tyler and Cutler, 2009; Ordin and Nespor, 2013). Thus, regarding duration, language-universal factors may play a bigger role for speech segmentation than language-specific word stress patterns (for a discussion, see Crowhurst, 2016).

The Relative Strengths of Different Prosodic Cues

Turning to the relative strengths of the different prosodic cues, our study allows a precise quantitative evaluation of the effect of pauses, duration, and pitch manipulations relative to a common statistical baseline. Overall, pauses between words provided the most helpful cues for speech segmentation. Especially in Experiment 3, which involved a strong memory component and was thus the most challenging, pauses outranked most other cues in effect (see non-overlapping CIs for words in all but one cue in **Figure 9**). This may be because pauses involve an immediate and very salient decrease in signal amplitude, relative to the other cues we tested (Fletcher, 2010). Additionally, pauses should provide nearly unambiguous signals for word boundaries because, in real speech, pauses almost exclusively occur at word boundaries and are not as flexibly distributed as changes in duration or pitch (Trainor and Adams, 2000; Fletcher, 2010; Matzinger et al., 2020).

Besides pauses, durational cues proved to be highly relevant cues for speech segmentation. In Experiment 1, final lengthening aided segmentation roughly as much as pauses (see **Figure 4**, and overlapping CIs in **Figure 5**), and final shortening was powerful enough to override statistical cues entirely.

In contrast, and perhaps surprisingly given previous results, pitch cues did not have very strong effects: in Experiment 3, performance when word-final pitch was increased was higher than when it was decreased, but performance based on modified pitch did not differ significantly from baseline performance in any of our experiments (see overlapping CIs in **Figures 4**, **8**). Thus, neither final pitch increase nor pitch decrease greatly affected speech segmentation. This result is concordant with the null effects of final pitch increase in German, Italian, Spanish and English (Toro et al., 2009; Tyler and Cutler, 2009; Ordin and Nespor, 2016), but contrasts with the facilitating effect of final pitch increase in French (most likely due to language-specific stress patterns of French: Tyler and Cutler, 2009). This may be because the pitch cues were perceived as neither converging or conflicting with the statistical cues, or because any perceived cue conflict was not strong enough to override the ever-present statistical cues. Investigations with more languages (especially tonal languages) employing a wider range of pitch changes would help resolve the role of pitch in word segmentation in adults.

Overall, our results tentatively suggest that durational cues are more relevant for speech segmentation than pitch cues (cf. Männel and Friederici, 2016), and that boundary cues of pauses and length might play a bigger role for segmentation than language-specific stress patterns, at least for the manipulation sizes employed here.

One possible reason for the primacy of durational information is that durational changes are language-universally more reliable cues for domain-finality than are pitch changes (Vaissière, 1983; Tyler and Cutler, 2009; Fletcher, 2010). In contrast, pitch changes often map onto language-specific word stress patterns (Tyler and Cutler, 2009; Ordin and Nespor, 2013, 2016; Ordin et al., 2017). In real speech, word stress can vary more than domain finality at the phrasal level (Ordin and Nespor, 2016), e.g., due to loan words with non-typical stress patterns (Broselow, 2009; Speyer, 2009; Andersson et al., 2017). Thus, pitch cues in natural speech may be employed more flexibly and variably than durational cues, making them less informative for speech segmentation. Although pitch changes also occur domain finally in real speech (e.g., final pitch decrease in declarative sentences or final pitch increase in yes-no questions; Vaissière, 1983), they may not have the same perceptual salience as durational cues. This may also explain why, overall, we found no clear differences between two opposing pitch changes: pitch decrease and pitch increase.

Robustness of the Results and Sensitivity to the Testing Environment

Although our results were consistent overall in the three studies, somewhat surprisingly, the effects described above unfolded most clearly in the most informal Experiment 1, less clearly in laboratory Experiment 2 and least clearly in Experiment 3. Experiment 3 was probably closest to most previous artificial language learning experiments in the literature. Thus, despite their overall consistency, our results were sensitive to differences in the experimental environment and the overall testing paradigm (for an overview of the methodological differences between the experiments and a summary of the results, see **Table 3**). Indeed, only in Experiment 1 did we replicate the finding that statistical cues alone suffice for successful speech segmentation, despite such effects being well-attested in the literature (e.g., Saffran et al., 1996a; Aslin et al., 1998).

A potential explanation for why the effects unfolded most clearly in Experiment 1 might be that in Experiments 2 and 3, in which participants were tested on more than one condition, participants were less focused in later stages of the experiment, which may be reflected in their overall segmentation scores. Also, in Experiments 2 and 3, participants may have inferred a segmentation rule (such as "the modified syllable is always the initial/medial/final syllable of the word") early on that they then transferred to later conditions. Depending on the rule they formed, this could either facilitate or impair segmentation in subsequent conditions. Because of the randomized order of the conditions and the null effect of the factor *order of the conditions* in our models, it is unlikely that there were consistent biases in a specific direction, but overall, cross-condition interference may have led to fuzzier results in Experiments 2 and 3.

The fact that Experiment 1 used syllables recorded by a native speaker of English does not seem to have influenced the overall pattern of results. If sub-phonemic cues in the English syllables had confused the listeners, results in Experiment 1 would have been expected to be fuzzier. In contrast, listeners may even have applied language-universal segmentation strategies such as final lengthening more consistently in Experiment 1 because they may have recognized that the syllables were not German and in turn reasoned that German-specific segmentation strategies may not be reliable in this case (cf. Quam and Creel, 2017).

Response Strategies in the Three Different Experiments

The slightly different setups in the three experiments appear to have led to different response strategies of the participants. In Experiment 2, participants made their choices most conservatively, meaning that overall they selected fewer test items as "words." This led to many misses of words and in general a lower performance on the identification of statistical words than on the rejection of statistical part-words. One potential reason is that, when participants tentatively identified a word, they then waited until this word reoccurred in the speech stream before confirming their choice and circling the

TABLE 3 | Summary of the methodological details and main results of the three experiments.

	Methods				Main results			
	Setting	**Design**	**Modality of test stimuli**	**Language of stimuli speaker**	**Baseline**	**Pauses**	**Durational cues**	**Pitch cues**
Exp. 1	Natural	Between-subjects	Visual	English	Successful segmentation	Improve segmentation	Lengthening improves & shortening hinders segmentation	No effect compared to baseline
Exp. 2	Lab	Within-subjects	Visual	German	No successful segmentation	Improve segmentation	Lengthening: successful segmentation, no improvement compared to baseline Shortening hinders segmentation	Not tested
Exp. 3	Lab	Within-subjects	Auditory	German	No successful segmentation	Improve segmentation	No effect compared to baseline; tendency: lengthening improves & shortening hinders segmentation	No effect compared to baseline; tendency: higher pitch improves & lower pitch hinders segmentation

item on the test sheet. They might not have had adequate time using this conservative strategy to identify all words while the speech stream was playing. However, participants did not exhibit this behavior in Experiment 1, although there they only had half the stimulus exposure as in Experiment 2. It is also possible that in the laboratory environment in Experiment 2, participants used explicit learning mechanisms, were more nervous or more concerned about doing well in the task and therefore answered more carefully and conservatively (cf. Parsons, 1974; Wickstrom and Bendix, 2000; Chiesa and Hobbs, 2008), whereas the informal environment in Experiment 1 triggered more implicit learning mechanisms, and elicited more immediate and thus perhaps more natural and spontaneous responses.

In contrast to Experiment 2, Experiment 3 was also in the laboratory, but this time had a pronounced auditory memory component. Here, participants overall chose very many items as "words." This led to many false alarms and in general a poor performance on statistical part-words compared to statistical words. Potentially, participants may have distrusted their memory and selected many items that sounded similar to those that they remembered. Overall, the differences between these two experiments suggests that when the task involves a pronounced memory component (similar to real language learning), speech segmentation becomes more challenging. On the one hand, the additional cognitive load of having to remember some segmented words might have made it more challenging for participants to extract later items from the stream. On the other hand, participants might have segmented many words correctly while listening to the speech stream, but then forgotten them later during the test phase. In any case, although participants performed worse in Experiment 3, their overall response patterns differed between the different cues as in the previous two experiments. The slightly different response patterns observed in our three experiments suggest that future speech segmentation studies should pay careful attention to such seemingly minor experimental differences, and it may be valuable to increase ecological validity by designing tasks that resemble real-life language learning more closely.

CONCLUSION AND OUTLOOK

In sum, our study provides new insights into how different prosodic cues aid or hinder statistics-based speech segmentation in native German-speaking adults. Because our study only manipulated word-final syllables, it would be interesting to replicate our study using the same manipulations, but on the initial or medial syllables of trisyllabic words (as done in Saffran et al., 1996b; Toro-Soto et al., 2007; Toro et al., 2009; Tyler and Cutler, 2009; Ordin and Nespor, 2016; Frost et al., 2017; Ordin et al., 2017; but these studies did not test opposing cues in direct comparison). Such a research program would provide a more comprehensive overview of the influence of the different individual cues in different locations. Our results make clear predictions for follow up-experiments with cues implemented on the medial syllables, especially for shortening cues. Since our study indicates that shortening word-medially sounds "most natural," even when this conflicts with statistical cues, medial shortening cues that match the statistical cues should lead to a higher segmentation performance (at least for German speakers). On the contrary, medial lengthening should hinder statistics-based segmentation performance.

Further tests manipulating word-initial cues would also be interesting with regard to the iambic-trochaic law (= ITL; Bolton, 1894; Hayes, 1995; Hay and Saffran, 2012; De la Mora et al., 2013; Abboub et al., 2016). Our study provides further evidence that, considering durational cues, the ITL generalizes from disyllabic to trisyllabic stimuli, namely that lengthened syllables are interpreted as word-final and lead to anapestic grouping (cf. Saffran et al., 1996b; Trainor and Adams, 2000; Tyler and Cutler, 2009; Frost et al., 2017). However, our null results regarding pitch modifications did not provide clear evidence regarding whether the ITL also generalizes to trisyllabic stimuli, leading to dactylic grouping. According to the ITL, higher pitched syllables are grouped sequence initially, so we predicted for our study that final pitch increase should hinder speech segmentation performance. However, this was not evident in our data. Thus, a variant of our experiment manipulating word-initial pitch would test more directly whether the ITL also transfers to trisyllabic stimuli for pitch modifications. If initial pitch increase indeed turned out to lead to dactylic grouping, this, combined with our finding about anapestic grouping of lengthened syllables, would point toward an "anapest-dactyl law" for trisyllabic stimuli, directly analogous to the ITL for bisyllabic stimuli.

Overall, we showed that different prosodic cues, namely pauses after the final syllables of trisyllabic statistical words, and durational and pitch cues on the final syllables of such words, had differing effects on speech segmentation. More specifically, pauses were most salient, duration changes also significant, and pitch changes showed little or no effect. Our findings are consistent with previous results indicating that when in conflict, prosodic cues can override statistical cues (e.g., Johnson and Jusczyk, 2001; Thiessen and Saffran, 2003; Johnson and Seidl, 2009). In addition, we found that changes in a single prosodic cue—duration—were enough to achieve such an override. Because we tested opposing cues—lengthening vs. shortening and pitch increase vs. decrease—in direct comparison, we were able to show that overall, durational cues played a more important and consistent role than pitch cues. These results contribute to a better understanding of which specific acoustic factors are most salient for listeners as they solve the challenge of speech segmentation.

Like most previous experimental work, our study tested speech segmentation in an artificial language with highly controlled and simplified stimuli and cue manipulations (although our study did use modified natural speech, rather than synthesized speech). This control and simplification has the virtue that the effects can be attributed to specific individual cues, but also raises the problem of how well these findings will translate to the segmentation of natural languages, where cues hardly ever occur in isolation and are more complex (Johnson and Seidl, 2009; Johnson and Tyler, 2010; Erickson and Thiessen, 2015). Although the full complexity of natural languages is hard to model in speech segmentation experiments in a controlled way, one step toward natural language conditions is to test durational, pitch and pause cues in combination, either converging or conflicting (like Ordin and Nespor, 2016 did for lengthening and pitch increase cues), but additionally adding pause, shortening and pitch decrease cues. This can shed light on whether the effects of cue changes are simply additive or if they interact in more complex ways when occurring in combination. Further factors that could move this research field toward natural languages include adding cues such as co-articulation cues (as e.g., in Johnson and Jusczyk, 2001), adding cues distinguishing between different boundary strengths (as e.g., in Sohail and Johnson, 2016), modifying the surrounding prosodic context (Morrill et al., 2014a,b, 2015), using words of different lengths and syllable structures (as e.g., in Johnson and Tyler, 2010), or incorporating prior lexical knowledge (as e.g., in Mattys et al., 2005), cues about syntactic structure (as e.g., Mueller et al., 2018), or even visual facial expression cues (as e.g., in Mitchel and Weiss, 2014). These all could be integrated in segmentation experiments with contrasting lengthening and pitch cues.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found here: https://osf.io/xtf6k/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the ethics board of the University of Vienna. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

TM: conceptualization, methodology, software, investigation, and writing—original draft preparation. NR: supervision and writing—review & editing. WTF: conceptualization, methodology, supervision, and writing—review & editing. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Austrian Science Fund (FWF) DK Grant Cognition & Communication [grant number: #W1262-B29] to WTF.

ACKNOWLEDGMENTS

We thank Roger Mundry for his invaluable help with data analysis and R scripts, Magdalena Schwarz for help with data collection, Felix Haiduk for helpful discussions, and two reviewers for their many helpful suggestions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.622042/full#supplementary-material

REFERENCES

Abboub, N., Boll-Avetisyan, N., Bhatara, A., Höhle, B., and Nazzi, T. (2016). An exploration of rhythmic grouping of speech sequences by french- and german-learning infants. *Front. Hum. Neurosci.* 10:292. doi: 10.3389/fnhum.2016.00292

Andersson, S., Sayeed, O., and Vaux, B. (2017). *The Phonology of Language Contact.* In Oxford Handbooks, 1–33. doi: 10.1093/oxfordhb/9780199935345.013.55

Aslin, R., Saffran, J., and Newport, E. (1998). Computation of conditional probability statistics by human infants. *Psychol. Sci.* 9, 321–324. doi: 10.1111/1467-9280.00063

Baayen, R. H. (2008). *Analyzing Linguistic Data*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511801686

Bagou, O., Fougeron, C., and Frauenfelder, U. H. (2002). "Contribution of prosody to the segmentation and storage of "words" in the acquisition of a new mini-language," in *Proceedings of the Speech Prosody 2002 Conference* (Aix-en-Provence), 159–162.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B. M., and Walker, S. C. (2015). Fitting linear mixed-effects models using Lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Bhatara, A., Boll-Avetisyan, N., Unger, A., Nazzi, T., and Höhle, B. (2013). Native language affects rhythmic grouping of speech. *J. Acoust. Soc. Am.* 134, 3828–3843. doi: 10.1121/1.4823848

Bion, R. A. H., Benavides-Varela, S., and Nespor, M. (2011). Acoustic markers of prominence influence infants' and adults' segmentation of speech sequences. *Lang. Speech* 54, 123–140. doi: 10.1177/0023830910388018

Boersma, P., and Weenik, D. (2017). *Praat: Doing Phonetics by Computer*. Available online at: http://www.praat.org/

Bolton, T. L. (1894). Rhythm. *Am. J. Psychol.* 6, 145–238. doi: 10.2307/1410948

Broselow, E. (2009). *Stress Adaptation in Loanword Phonology In Phonology in Perception, edited by Paul Boersma and Silke Hamann, 191–234*. Berlin: De Gruyter Mouton.

Brown, G. S., and White, K. G. (2005). The optimal correction for estimating extreme discriminability. *Behav. Res. Methods* 37, 436–449. doi: 10.3758/BF03192712

Chiesa, M., and Hobbs, S. (2008). Making sense of social research: how useful is the hawthorne effect? *Eur. J. Soc. Psychol.* 38, 67–74. doi: 10.1002/ejsp.401

Christophe, A., Peperkamp, S., Pallier, C., Block, E., and Mehler, J. (2004). Phonological phrase boundaries constrain lexical access i. adult data. *J. Mem. Lang.* 51, 523–547. doi: 10.1016/j.jml.2004.07.001

Cole, R. A., Jakimik, J., and Cooper, W. E. (1980). Segmenting speech into words. *J. Acoust. Soc. Am.* 67, 1323–1332. doi: 10.1121/1.384185

Crowhurst, M. (2016). Iambic-Trochaic Law Effects among Native Speakers of Spanish and English. *Lab. Phonol.* 7:12. doi: 10.5334/labphon.42

Cruttenden, A. (1986). *Intonation*. *Studies in English Literature*. Cambridge: Cambridge University Press.

Cumming, G. (2012). *Understanding The New Statistics*. New York: Routledge. doi: 10.4324/9780203807002

Cumming, G. (2014). The new statistics: why and how. *Psychol. Sci.* 25, 7–29. doi: 10.1177/0956797613504966

Cumming, G., and Finch, S. (2005). Inference by eye confidence intervals and how to read pictures of data. *Am. Psychol.* 60, 170–180. doi: 10.1037/0003-066X.60.2.170

Cutler, A. (1990). "Exploiting prosodic probabilities in speech segmentation," in *Cognitive Models of Speech Processing*, eds G. T. M. Altmann (Cambridge, MA: MIT Press), 105–21.

Cutler, A. (1991). Linguistic rhythm and speech segmentation. *Music Lang. Speech Brain* 157–66. doi: 10.1007/978-1-349-12670-5_14

Cutler, A. (2005). "Lexical stress," in *The Handbook of Speech Perception*, eds David B. Pisoni and Robert E. Remez (Malden, MA: Blackwell), 264–89. doi: 10.1002/9780470757024.ch11

Cutler, A., Dahan, D., and Van Donselaar, W. (1997). Prosody in the comprehension of spoken language : a literature review. *Lang. Speech* 40, 141–202. doi: 10.1177/002383099704000203

Cutler, A., Mehler, J., Norris, D., and Segui, J. (1992). The monolingual nature of speech by bilinguals. *Cogn. Psychol.* 24, 381–410. doi: 10.1016/0010-0285(92)90012-Q

Dahan, D., and Magnuson, J. S. (2006). "Spoken word recognition," in *Handbook of Psycholinguistics*, eds Matthew J. Traxler and Morton A. Gernsbacher (Cambridge, Mass.: Academic Press), 249–83. doi: 10.1016/B978-012369374-7/50009-2

De la Mora, D. M., Nespor, M., and Toro, J. M. (2013). Do Humans and nonhuman animals share the grouping principles of the iambic – trochaic law? *Atten. Percept. Psychophys* 75, 92–100. doi: 10.3758/s13414-012-0371-3

Delattre, P. (1965). *Comparing the Phonetic Features of English, German, Spanish and French*. Heidelberg: J. Groos. doi: 10.1515/iral.1964.2.1.155

Dilley, L. C., and McAuley, J. D. (2008). Distal prosodic context affects word segmentation and lexical processing. *J. Mem. Lang.* 59, 294–311. doi: 10.1016/j.jml.2008.06.006

Dobson, A. J. (2002). *An Introduction to Generalized Linear Models*. Boca Raton: Chapman & Hall. doi: 10.1201/9781420057683

Dogil, G., and Williams, B. (1999). "The phonetic manifestation of word stress," in *Word Prosodic Systems in the Languages of Europe*, eds Harry van der Hulst (Berlin: Mouton de Gruyter), 273–310. doi: 10.1515/9783110197082.1.273

Domahs, U., Plag, I., and Carroll, R. (2014). Word stress assignment in german, english and dutch: quantity-sensitivity and extrametricality revisited. *J. Comp. Germ. Lingu.* 17, 59–96. doi: 10.1007/s10828-014-9063-9

El Zarka, D., Schuppler, B., Lozo, C., Eibler, W., and Wurzwallner, P. (2017). "Acoustic correlates of stress and accent in standard Austrian German," in *Phonetik in Und Über Österreich*, eds S. Moosmüller, C. Schmid, and M. Sellner (Vienna: Verlag der Österreichischen Akademie der Wissenschaften), 15–44. doi: 10.2307/j.ctt1v2xvhh.5

Endress, A. D., and Hauser, M. D. (2010). Word segmentation with universal prosodic cues. *Cogn. Psychol.* 61, 177–199. doi: 10.1016/j.cogpsych.2010.05.001

Endress, A. D., and Mehler, J. (2009). The surprising power of statistical learning: when fragment knowledge leads to false memories of unheard words. *J. Mem. Lang.* 60, 351–367. doi: 10.1016/j.jml.2008.10.003

Erickson, L. C., and Thiessen, E. D. (2015). Statistical learning of language: theory, validity, and predictions of a statistical learning account of language acquisition. *Dev. Rev.* 37, 66–108. doi: 10.1016/j.dr.2015.05.002

Ernestus, M., and Neijt, A. (2008). Word length and the location of primary word stress in Dutch, German, and English. *Linguistics* 46, 507–540. doi: 10.1515/LING.2008.017

Féry, C. (1998). German word stress in optimality theory. *J. Comp. Germ. Lingu.* 2, 101–142. doi: 10.1023/A:1009883701003

Féry, C., Hörnig, R., and Pahaut, S. (2011). "Correlates of phrasing in french and german from an experiment with semi-spontaneous speech," in *Intonational Phrasing in Romance and Germanic: Cross-Linguistic and Bilingual Studies*, eds C. Gabriel and C. Lleó (Hamburg: John Benjamins), 11–41. doi: 10.1075/hsm.10.03fer

Filippi, P., Gingras, B., and Fitch, W. T. (2014). Pitch enhancement facilitates word learning across visual contexts. *Front. Psychol.* 5, 1–8. doi: 10.3389/fpsyg.2014.01468

Fletcher, J. (2010). "The prosody of speech : timing and rhythm," in *The Handbook of Phonetic Sciences*, eds W. J. Hardcastle, J. Laver, and F. E. Gibbon, 2nd ed (Hoboken: Wiley-Blackwell), 523–602. doi: 10.1002/9781444317251.ch15

Friederici, A. D., and Männel, C. (2013). "Neural correlates of the development of speech perception and comprehension," in *The Oxford Handbook of Cognitive Neuroscience*, eds K. Ochsner and S. M. Kosslyn. Vol. 1 (Oxford: Oxford University Press), 1–36. doi: 10.1093/oxfordhb/9780199988693.013.0009

Frost, R. L. A., Monaghan, P., and Tatsumi, T. (2017). Domain-general mechanisms for speech segmentation: the role of duration information in language learning. *J. Exp. Psychol.* 43, 466–476. doi: 10.1037/xhp0000325

Giegerich, H. J. (1985). *Metrical Phonology and Phonological Structure: German and English. Vol. 43. Cambridge Studies in Linguistics*. Cambridge: Cambridge University Press.

Gout, A., Christophe, A., and Morgan, J. L. (2004). Phonological phrase boundaries constrain lexical access ii. infant data. *J. Mem. Lan.g* 51, 548–567. doi: 10.1016/j.jml.2004.07.002

Green, D. M., and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. New York, NY: John Wiley.

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d'. *Behav. Res. Methods,Instru. Comp.* 27, 46–51. doi: 10.3758/BF03203619

Hay, J. S. F., and Diehl, R. L. (2007). Perception of rhythmic grouping: testing the iambic/trochaic law. *Perecep. Psychophys.* 69, 113–122. doi: 10.3758/BF03194458

Hay, J. S. F., and Saffran, J. R. (2012). Rhythmic grouping biases constrain infant statistical learning. *Infancy* 17, 610–641. doi: 10.1111/j.1532-7078.2011.00110.x

Hayes, B. (1995). *Metrical Stress Theory: Principles and Case Studies*. Chicago: The University of Chicago Press.

Hayes, J. R., and Clark, H. H. (1970). "Experiments in the segmentation of an artificial speech analog," in *Cognition and the Development of Language*, ed J. R. Hayes (New York, NY: Wiley), 221–234.

Hirst, D., and Di Cristo, A. (eds). (1998). *Intonation Systems - a Survey of Twenty Languages. Intonation Systems: A Survey of Twenty Languages*. Cambridge: Cambridge University Press.

Isachenko, A. V., and Schädlich, H. J. (1966). Untersuchungen über die deutsche satzintonation. *Studia Grammatica* 7, 7–64.

Iversen, J. R., Patel, A. D., and Ohgushi, K. (2008). perception of rhythmic grouping depends on auditory experience. *J. Acoust. Soc. Am.* 124, 2263–2271. doi: 10.1121/1.2973189

Johnson, E. K. (2008). Infants use prosodically conditioned acoustic-phonetic cues to extract words from speech. *J. Acoust. Soc. Am.* 123, EL144–48. doi: 10.1121/1.2908407

Johnson, E. K. (2012). "Bootstrapping language : are infant statisticians up to the job?" in *Statistical Learning and Language Acquisition*, eds P. Rebuschat and J. Williams (Berlin: Mouton de Gruyter), 55–90.

Johnson, E. K. (2016). Constructing a proto-lexicon : an integrative view of infant language development. *Ann. Rev. Ling.* 2, 391–412. doi: 10.1146/annurev-linguistics-011415-040616

Johnson, E. K., and Jusczyk, P. W. (2001). Word segmentation by 8-month-olds : when speech cues count more than statistics. *J. Mem. Lang.* 44, 548–567. doi: 10.1006/jmla.2000.2755

Johnson, E. K., Lahey, M., Ernestus, M., and Cutler, A. (2013). A multimodal corpus of speech to infant and adult listeners. *J. Acoust. Soc. Am.* 134, EL534–40. doi: 10.1121/1.4828977

Johnson, E. K., Seidl, A., and Tyler, M. D. (2014). The edge factor in early word segmentation : utterance-level prosody enables word form extraction by 6-month-olds. *PLoS ONE* 9:e83546. doi: 10.1371/journal.pone.0083546

Johnson, E. K., and Seidl, A. H. (2009). At 11 months, prosody still outranks statistics. *Dev. Sci.* 12, 131–141. doi: 10.1111/j.1467-7687.2008.00740.x

Johnson, E. K., and Tyler, M. D. (2010). Testing the limits of statistical learning for word segmentation. *Dev. Sci.* 13, 339–345. doi: 10.1111/j.1467-7687.2009.00886.x

Kahng, J. (2014). Exploring utterance and cognitive fluency of l1 and l2 english speakers: temporal measures and stimulated recall. *Lang. Learn.* 64, 809–854. doi: 10.1111/lang.12084

Kim, S., Broersma, M., and Cho, T. (2012). The use of prosodic cues in learning new words in an unfamiliar language. *Stud. Second Lang. Acquis.* 34, 415–444. doi: 10.1017/S0272263112000137

Kirby, S., Cornish, H., and Smith, K. (2008). Cumulative cultural evolution in the laboratory: an experimental approach to the origins of structure in human language. *Proc. Natl. Acad. Sci. U.S.A.* 105, 10681–10686. doi: 10.1073/pnas.0707835105

Klatt, D. H. (1975). Vowel lengthening is syntactically determined in connected discourse. *J. Phon.* 3, 129–140. doi: 10.1016/S0095-4470(19) 31360-9

Kohler, K. J. (2012). The perception of lexical stress in german: effects of segmental duration and vowel quality in different prosodic patterns. *Phonetica* 69, 68–93. doi: 10.1159/000342126

Kohler, K. J., and Rodgers, J. E. J. (2001). *Schwa Deletion in German Read and Spontaneous Speech. Spontaneous German Speech: Symbolic Structures and Gestural Dynamics*, 97–123. Available online at: http://www.ipds.uni-kiel.de/kjk/pub_exx/aipuk35/kkjr.pdf

Langus, A., Marchetto, E., Bion, R. A. H., and Nespor, M. (2012). Can prosody be used to discover hierarchical structure in continuous speech? *J. Mem. Lang.* 66, 285–306. doi: 10.1016/j.jml.2011.09.004

Macmillan, N. A., and Creelman, C. D. (2005). *Detection Theory: A User's Guide. 2nd ed.* Mahwah: Lawrence Erlbaum.

Macmillan, N. A., and Kaplan, H. L. (1985). Detection theory analysis of group data: estimating sensitivity from average hit and false-alarm rates. *Psychol. Bull.* 98, 185–199. doi: 10.1037/0033-2909.98.1.185

Männel, C., and Friederici, A. D. (2016). Neural correlates of prosodic boundary perception in german preschoolers: if pause is present, pitch can go. *Brain Res.* 1632, 27–33. doi: 10.1016/j.brainres.2015.12.009

Marimon Tarter, M. (2019). *Word Segmentation in German-Learning Infants and German-Speaking Adults: Prosodic and Statistical Cues*. Potsdam: University of Potsdam.

Mattys, S. L., Jusczyk, P. W., Luce, P. A., and Morgan, J. L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cogn. Psychol.* 38, 465–494. doi: 10.1006/cogp.1999.0721

Mattys, S. L., White, L., and Melhorn, J. F. (2005). Integration of multiple speech segmentation cues: a hierarchical framework. *J. Exp. Psychol. Gen.* 134, 477–500. doi: 10.1037/0096-3445.134.4.477

Matzinger, T., Ritt, N., and Fitch, W. T. (2020). Non-native speaker pause patterns closely correspond to those of native speakers at different speech rates. *PLoS ONE* 15:e0230710. doi: 10.1371/journal.pone.0230710

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models*. London: Chapman & Hall. doi: 10.1007/978-1-4899-3242-6

Mitchel, A. D., and Weiss, D. J. (2014). Visual speech segmentation: using facial cues to locate word boundaries in continuous speech. *Lang. Cogn. Neurosci.* 29, 771–780. doi: 10.1080/01690965.2013.791703

Morgan, J. L., and Saffran, J. R. (1995). Emerging integration of sequential and suprasegmental information in preverbal speech segmentation. *Child Dev.* 66, 911–936. doi: 10.2307/1131789

Morrill, T. H., Dilley, L. C., and McAuley, J. D. (2014a). Prosodic patterning in distal speech context: effects of list intonation and f0 downtrend on perception of proximal prosodic structure. *J. Phon.* 46, 68–85. doi: 10.1016/j.wocn.2014.06.001

Morrill, T. H., Dilley, L. C., McAuley, J. D., and Pitt, M. A. (2014b). Distal rhythm influences whether or not listeners hear a word in continuous speech: support for a perceptual grouping hypothesis. *Cognition* 131, 69–74. doi: 10.1016/j.cognition.2013.12.006

Morrill, T. H., McAuley, J. D., Dilley, L. C., Zdziarska, P. A., Jones, K. B., and Sanders, L. D. (2015). Distal prosody affects learning of novel words in an artificial language. *Psychonomic Bull. Rev.* 22, 815–823. doi: 10.3758/s13423-014-0733-z

Moulines, E., and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Commun.* 9, 453–467. doi: 10.1016/0167-6393(90)90021-Z

Mueller, J. L., Bahlmann, J., and Friederici, A. D. (2010). Learnability of embedded syntactic structures depends on prosodic cues. *Cogn. Sci.* 34, 338–349. doi: 10.1111/j.1551-6709.2009.01093.x

Mueller, J. L., Ten Cate, C., and Toro, J. M. (2018). A comparative perspective on the role of acoustic cues in detecting language structure. *Top. Cogn. Sci.* 12, 1–16. doi: 10.1111/tops.12373

Munafò, M. R., and Smith, G. D. (2018). Repeating experiments is not enough. *Nature* 553, 399–401. doi: 10.1038/d41586-018-01023-3

Nespor, M., Shukla, M., Van De Vijver, R., Avesani, C., Schraudolf, H., and Donati, H. (2008). Different phrasal prominence realizations in VO and OV languages. *Lingue e Linguaggio* 7, 139–167. doi: 10.1418/28093

Norris, D., and Cutler, A. (1988). The role of strong syllables in segmentation for lexical access. *J. Exp. Psychol.* 14, 113–121. doi: 10.1037/0096-1523.14.1.113

O'Brien, M. G., and Fagan, S. M. B. (2016). *German Phonetics and Phonology : Theory and Practice*. New Haven: Yale University Press.

Oller, D. K. (1973). The effect of position in utterance on speech segment duration in english. *J. Acoust. Soc. Am.* 54, 1235–1247. doi: 10.1121/1.1914393

Ordin, M., and Nespor, M. (2013). Transition probabilities and different levels of prominence in segmentation. *Lang. Learn.* 63, 800–834. doi: 10.1111/lang.12024

Ordin, M., and Nespor, M. (2016). Native language influence in the segmentation of a novel language. *Lang. Learn. Dev.* 12, 461–481. doi: 10.1080/15475441.2016.1154858

Ordin, M., Polyanskaya, L., Laka, I., and Nespor, M. (2017). Cross-linguistic differences in the use of durational cues for the segmentation of a novel language. *Mem. Cognit.* 45, 863–876. doi: 10.3758/s13421-017-0700-9

Palmer, S. D., and Mattys, S. L. (2016). Speech segmentation by statistical learning is supported by domain-general processes within working memory. *Q. J. Exp. Psychol.* 69, 2390–2401. doi: 10.1080/17470218.2015.1112825

Pamies Bertrán, A. (1999). Prosodic typology: on the dichotomy between stress-timed and syllable-timed languages. *Language Design* 2, 103–130.

Parsons, H. M. (1974). What happened at hawthorne? *Science* 183, 922–932. doi: 10.1126/science.183.4128.922

Peirce, J. (2007). PsychoPy: psychophysics software in python. *J. Neurosci.* 162, 8–13. doi: 10.1016/j.jneumeth.2006.11.017

Peña, M., Bonatti, L. L., Nespor, M., and Mehler, J. (2002). Signal-driven computations in speech processing. *Science* 298, 604–607. doi: 10.1126/science.1072901

Pierce, L. J., Genesee, F., Delcenserie, A., and Morgan, G. (2017). Variations in phonological working memory: linking early language experiences and language learning outcomes. *Appl. Psycholinguist* 38, 1265–1300. doi: 10.1017/S0142716417000236

Quam, C., and Creel, S. C. (2017). Mandarin-english bilinguals process lexical tones in newly learned words in accordance with the language context. *PLoS ONE* 12:e0169001. doi: 10.1371/journal.pone.0169001

R Development Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available online at: http://www.r-project.org/

Romberg, A. R., and Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdiscipl. Rev. Cogn. Sci.* 1, 906–914. doi: 10.1002/wcs.78

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996a). Statistical learning by 8-month-old infants. *Science* 274, 1926–1928. doi: 10.1126/science.274.5294.1926

Saffran, J. R., Elizabeth, K., Johnson, Richard, N., Aslin, and Elissa, L., Newport. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition* 70, 27–52. doi: 10.1016/S0010-0277(98)00075-4

Saffran, J. R., Newport, E. L., and Aslin, R. N. (1996b). Word segmentation : the role of distributional cues. *J. Mem. Lang.* 35, 606–621. doi: 10.1006/jmla.1996.0032

Schielzeth, H., and Forstmeier, W. (2009). Conclusions beyond support: overconfident estimates in mixed models. *Behav. Ecol.* 20, 416–420. doi: 10.1093/beheco/arn145

Seidl, A. (2007). Infants' use and weighting of prosodic cues in clause segmentation. *J. Mem. Lang.* 57, 24–48. doi: 10.1016/j.jml.2006.10.004

Shrout, P. E., and Rodgers, J. L. (2018). Psychology, science, and knowledge construction: broadening perspectives from the replication crisis. *Annu. Rev. Psychol.* 69, 487–510. doi: 10.1146/annurev-psych-122216-011845

Shukla, M., Nespor, M., and Mehler, J. (2007). An interaction between prosody and statistics in the segmentation of fluent speech. *Cogn. Psychol.* 54, 1–32. doi: 10.1016/j.cogpsych.2006.04.002

Sohail, J., and Johnson, E. K. (2016). How transitional probabilities and the edge effect contribute to listeners' phonological bootstrapping success. *Langu. Learn. Dev.* 12, 105–115. doi: 10.1080/15475441.2015.1073153

Song, J. Y., Demuth, K., and Morgan, J. (2010). Effects of the acoustic properties of infant-directed speech on infant word recognition. *J. Acoust. Soc. Am.* 128, 389–400. doi: 10.1121/1.3419786

Speyer, A. (2009). On the change of word stress in the history of German. *Beiträge zur Geschichte der deutschen Sprache Literatur* 131, 413–441. doi: 10.1515/bgsl.2009.051

Stanislaw, H., and Todorov, N. (1999). Calculation of signal detection theory measures. *Behav. Res. Methods Instrum.* 31, 137–149. doi: 10.3758/BF03207704

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cogn. Psychol.* 50, 86–132. doi: 10.1016/j.cogpsych.2004.06.001

Swingley, D. (2009). Onsets and codas in 1.5-year-olds' word recognition. *J. Mem. Lang.* 60, 252–269. doi: 10.1016/j.jml.2008.11.003

Thiessen, E. D., and Saffran, J. R. (2003). When cues collide: use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Dev. Psychol.* 39, 706–716. doi: 10.1037/0012-1649.39.4.706

Thiessen, E. D., and Saffran, J. R. (2007). Learning to learn: infants' acquisition of stress-based strategies for word segmentation. *Lang. Lear. Dev.* 3, 73–100. doi: 10.1207/s15473341lld0301_3

Toro, J. M., Sebastián-Gallés, N., and Mattys, S. L. (2009). The role of perceptual salience during the segmentation of connected speech. *Europ. J. Cogn. Psych.* 21, 786–800. doi: 10.1080/09541440802405584

Toro, J. M., Sinnett, S., and Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition* 97, 25–34. doi: 10.1016/j.cognition.2005.01.006

Toro-Soto, J. M., Rodríguez-Fornells, A., and Sebastián-Gallés, N. (2007). Stress placement and word segmentation by spanish speakers. *Psicológica* 4, 167–176. Available online at: https://www.redalyc.org/pdf/169/16928204.pdf

Trainor, L. J., and Adams, B. (2000). Infants' and adults' use of duration and intensity cues in the segmentation of tone patterns. *Percept. Psychophys* 62, 333–340. doi: 10.3758/BF03205553

Tyler, M. D., and Cutler, A. (2009). Cross-language differences in cue use for speech segmentation. *J. Acoust. Soc. Am.* 126, 367–376. doi: 10.1121/1.3129127

Vaissière, J. (1983). "Language-independent prosodic features," in *Springer Series in Language and Communication 14: Prosody: Models and Measurements,* eds A. Cutler and D. R. Ladd (Hamburg: Springer), 53–66. doi: 10.1007/978-3-642-69103-4_5

Wedel, A., Ussishkin, A., and King, A. (2019). Incremental word processing influences the evolution of phonotactic patterns. *Folia Lingu.* 40, 231–248. doi: 10.1515/flih-2019-0011

Wen, Z. (2016). *Working Memory and Second Language Learning: Towards an Integrated Approach*. Bristol: Channel View Publications. doi: 10.21832/9781783095735

White, L., Benavides-Varela, S., and Mády, K. (2020). Are initial-consonant lengthening and final-vowel lengthening both universal word segmentation cues? *J. Phon.* 81:100982. doi: 10.1016/j.wocn.2020.100982

Wickstrom, G., and Bendix, T. (2000). The ″hawthorne effect″ - what did the original hawthorne studies actually show? *Scand. J. Work Environ. Health* 26, 363–367. doi: 10.5271/sjweh.555

Woodrow, H. (1909). A quantitative study of rhythm: the effect of variations in intensity, rate and duration. *Arch. Psychol.* 14, 1–66.

Zellner, B. (1994). "Pauses and the temporal structure of speech," in *Fundamentals of Speech Synthesis and Speech Recognition*, eds E. Keller (Chichester: John Wiley), 41–62.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The influence of different prosodic cues on word segmentation

Theresa Matzinger[1,2*], Nikolaus Ritt[1], W. Tecumseh Fitch[2,3*]

[1] Department of English, University of Vienna, Spitalgasse 2/Hof 8, 1090 Vienna;

[2] Department of Behavioral and Cognitive Biology, University of Vienna, Althanstraße 14, 1090 Vienna;

[3] Cognitive Science Hub, University of Vienna, Vienna, Austria.

Supplementary materials

Experiment 1

Table S1. Results of the logistic generalized mixed model exploring the effects of condition, stimulus type, and their interaction, on the correctness of responses in a decision task where participants had to decide on whether presented stimuli were words or not. The table reports estimated model coefficients, standard errors (SE), lower and upper confidence intervals for the estimates, χ^2-values of likelihood ratio tests, respective degrees of freedom (df) and p-values (P, significances in bold), as well as minimum and maximum estimates obtained after dropping individuals one at a time (an indicator of model stability).

Term	**Estimate**	**SE**	**Lower CI**	**Upper CI**	**χ^2**	**df**	**P**	**Min. estimate**	**Max. estimate**
Intercept	0.84	0.32	0.23	1.55	(1)	(1)	(1)	0.77	0.94
conditionPAU	2.87	0.62	1.76	4.33				2.74	3.30
conditionLEN	0.94	0.48	-0.01	1.86				0.82	1.10
conditionSHO	-2.83	0.49	-3.87	-1.87				-2.93	-2.76
conditionLOW	0.46	0.46	-0.49	1.41				0.35	0.59
conditionHIG	0.25	0.46	-0.71	1.18				0.13	0.37
partword_typePW1	0.13	0.40	-0.68	0.95				0.00	0.27
partword_typePW2	-0.56	0.32	-1.20	0.06				-0.66	-0.42
conditionPAU:partword_typePW1	-1.46	0.73	-3.14	0.04	63.129	10	**<0.001**	-1.84	-1.20
conditionLEN:partword_typePW1	0.27	0.61	-0.97	1.45				0.12	0.45
conditionSHO:partword_typePW1	3.31	0.60	2.08	4.63				3.16	3.53
conditionLOW:partword_typePW1	0.27	0.58	-0.92	1.45				0.10	0.48
conditionHIG:partword_typePW1	0.18	0.57	-1.04	1.31				0.04	0.39
conditionPAU:partword_typePW2	-0.64	0.66	-2.18	0.64				-0.94	-0.40
conditionLEN:partword_typePW2	1.34	0.55	0.25	2.51				1.14	1.57
conditionSHO:partword_typePW2	2.32	0.50	1.40	3.40				2.17	2.47

conditionLOW:partword_typePW2	0.14	0.47	-0.81	1.00				0.00	0.29
conditionHIG:partword_typePW2	0.32	0.47	-0.65	1.26				0.18	0.45

(1) Not shown because it allows very limited interpretation

Experiment 2

Table S2. Results of the logistic generalized mixed model exploring the effects of condition, stimulus type, their interaction, order of the conditions, and language on the correctness of responses in a decision task where participants had to decide on whether presented stimuli were words or not. The table reports estimated model coefficients, standard errors (SE), lower and upper confidence intervals for the estimates, χ^2-values of likelihood ratio tests, respective degrees of freedom (df) and p-values (P, near-significances in bold), as well as minimum and maximum estimates obtained when dropping individuals one at a time (being an indicator of model stability).

Term	**Estimate**	**SE**	**Lower CI**	**Upper CI**	**χ^2**	**df**	**P**	**Min. estimate**	**Max. estimate**
Intercept	-0.45	0.31	-1.05	0.15	(1)	(1)	(1)	-0.55	-0.35
conditionPAU	2.24	0.43	1.45	3.22				2.07	2.42
conditionLEN	1.12	0.34	0.50	1.84				1.01	1.29
conditionSHO	-0.77	0.32	-1.39	-0.18				-0.92	-0.61
partword_typePW1	1.69	0.40	0.94	2.52				1.55	1.90
partword_typePW2	1.42	0.42	0.63	2.27				1.27	1.62
z.condition_order_no (2)	-0.09	0.10	-0.27	0.10	0.954	1	0.329	-0.14	-0.06
languageLanguage2 (3)	-0.09	0.24	-0.55	0.39	1.725	3	0.631	-0.22	0.00
languageLanguage3	-0.10	0.24	-0.56	0.39				-0.21	0.04
languageLanguage4	0.17	0.25	-0.31	0.71				0.06	0.29
conditionPAU:partword_typePW1	-1.02	0.49	-2.00	-0.05	11.329	6	**0.079**	-1.21	-0.78
conditionLEN:partword_typePW1	-0.76	0.46	-1.66	0.17				-0.92	-0.61
conditionSHO:partword_typePW1	0.30	0.53	-0.72	1.30				0.05	0.50
conditionPAU:partword_typePW2	-1.07	0.47	-2.07	-0.16				-1.26	-0.87
conditionLEN:partword_typePW2	-0.31	0.49	-1.28	0.73				-0.61	-0.15
conditionSHO:partword_typePW2	0.16	0.42	-0.68	0.97				0.02	0.30

(1) Not shown because it allows very limited interpretation

(2) z-transformed mean and standard deviation of the original variable were 1.50 and 1.12 respectively

(3) *language* was manually dummy coded with Language1 being the reference category, and then centered

Experiment 3

Table S3. Results of the logistic generalized mixed model exploring the effects of condition, stimulus type, their interaction, order of the conditions, trial number, and language on the correctness of responses in a decision task where participants had to decide on whether presented stimuli were words or not. The table reports estimated model coefficients, standard errors (SE), lower and upper confidence intervals for the estimates, χ^2-values of likelihood ratio tests, respective degrees of freedom (df) and p-values (P, significances in bold), as well as minimum and maximum estimates obtained when dropping individuals one at a time (being an indicator of model stability).

Term	**Estimate**	**SE**	**Lower CI**	**Upper CI**	**χ^2**	**df**	**P**	**Min. estimate**	**Max. estimate**
Intercept	0.97	0.19	0.60	1.39	(1)	(1)	(1)	0.92	1.06
conditionPAU	1.29	0.34	0.61	1.95				1.18	1.45
conditionLEN	0.30	0.38	-0.39	1.08				0.16	0.43
conditionSHO	-0.87	0.33	-1.50	-0.21				-1.01	-0.71
conditionLOW	-1.28	0.33	-1.94	-0.68				-1.39	-1.14
conditionHIG	0.00	0.32	-0.60	0.63				-0.16	0.20
partword_typePW1	-1.79	0.27	-2.38	-1.31				-1.93	-1.70
partword_typePW2	-1.52	0.25	-2.02	-1.06				-1.68	-1.43
z.condition_order_no (2)	0.02	0.06	-0.10	0.15	0.159	1	0.692	-0.01	0.04
z.trial_no (3)	-0.01	0.05	-0.11	0.10	0.014	1	0.907	-0.02	0.02
languageLanguage2 (4)	-0.12	0.18	-0.48	0.22	4.013	3	0.260	-0.19	-0.07
languageLanguage3	0.20	0.18	-0.14	0.58				0.08	0.29
languageLanguage4	0.15	0.17	-0.19	0.48				0.07	0.22
conditionPAU:partword_typePW1	-0.15	0.40	-0.99	0.68	31.963	10	**< 0.001**	-0.26	0.03
conditionLEN:partword_typePW1	-0.51	0.46	-1.54	0.33				-0.70	-0.35
conditionSHO:partword_typePW1	1.61	0.56	0.54	2.71				1.33	1.87
conditionLOW:partword_typePW1	1.85	0.47	0.89	2.79				1.67	2.01
conditionHIG:partword_typePW1	0.57	0.47	-0.33	1.50				0.33	0.76
conditionPAU:partword_typePW2	-0.38	0.39	-1.13	0.33				-0.51	-0.19
conditionLEN:partword_typePW2	-0.31	0.44	-1.20	0.53				-0.49	-0.16
conditionSHO:partword_typePW2	0.61	0.44	-0.31	1.47				0.41	0.80
conditionLOW:partword_typePW2	1.67	0.55	0.60	2.78				1.44	1.91
conditionHIG:partword_typePW2	0.51	0.42	-0.27	1.34				0.17	0.75

(1) Not shown because it allows very limited interpretation

(2) z-transformed, mean and sd of the original variable were 1.50 and 1.12, respectively
(3) z-transformed, mean and sd of the original variable were 5.50 and 3.45, respectively
(4) *language* was manually dummy coded with Language1 being the reference category, and then centered

CHAPTER 3

Aesthetic appeal of prosodic patterns influences their segmentation from a speech stream

This chapter is under review in the *Journal of Language Evolution*.

Citation:

Matzinger T, Specker E, Ritt N & Fitch WT. Aesthetic appeal of prosodic patterns influences their segmentation from a speech stream (*under review*)

Aesthetic appeal of prosodic patterns influences their segmentation from a speech stream

Theresa Matzinger[1,2*], Eva Specker[3], Nikolaus Ritt[1], W. Tecumseh Fitch[2]

[1] Department of English, University of Vienna, Austria

[2] Department of Behavioral and Cognitive Biology, University of Vienna, Austria

[3] Department of Cognition, Emotion, and Methods in Psychology, University of Vienna, Austria

*** Correspondence:**
Theresa Matzinger
theresa.matzinger@univie.ac.at

Keywords: aesthetics; prosody; speech segmentation; artificial language learning; syllable duration

Abstract

Could the aesthetic appeal of linguistic features influence their learnability, and in turn their long-term stability during language change? Focusing on prosodic patterns, we investigated the baseline assumption that linguistic features like stress patterns affect aesthetic appeal. Listeners' ratings of isochronous words and words with initially, medially or finally lengthened or shortened syllables revealed that, indeed, different prosodic patterns differed in their aesthetic appeal in terms of 'liking', 'beauty' and 'naturalness'. Furthermore, aesthetic ratings of these prosodic patterns corresponded to their effectiveness for speech segmentation in previous experiments, suggesting a potential connection between aesthetics and language processing. This research opens up new avenues for future research on the role of aesthetic preferences in language acquisition and language change.

Keywords: aesthetics; language change; prosody; speech segmentation; artificial language learning; syllable duration

Introduction

All human cultures appreciate beauty, and humans readily evaluate visual, verbal, or musical stimuli in terms of aesthetic appeal (Nadal & Vartanian, 2019; Rastall, 2008). Traditionally, aesthetics has been investigated in the visual domain, but more recently, research has been extended to the aesthetic appeal of other sensory domains including music, poetry, and literature (Verheyen, 2015). While research on the aesthetic appeal of linguistic features (such as Rastall, 2008) is still comparatively scarce, linguistic stimuli have often been investigated with regard to their processing fluency, emotional value, valence or arousal (Foolen, Lüdtke, Racine, & Zlatev, 2012; Forster, 2020; Laham, Koval, & Alter, 2012; Paulmann, Bleichner, & Kotz, 2013; Warriner, Kuperman, & Brysbaert, 2013) – features closely linked to aesthetic appeal (Leder, Ring, & Dressler, 2013; Reber, Schwarz, & Winkielman, 2004; Shibles, 1995). These connections make focused investigations of the aesthetic appeal of linguistic features worthwhile (Keller, 1994).

For example, aesthetic preferences regarding prosodic patterns play a prominent role in the appeal of poetry (Obermeier et al., 2016). If such preferences also apply in the perception of spontaneous everyday speech, they may affect language learning and, indirectly, also language change. A plausible hypothesis is that aesthetically appealing linguistic features are memorized more easily (Kousta, Vinson, & Vigliocco, 2009; Reber et al., 2004) and used more frequently, and therefore transmitted more successfully across speaker generations than less aesthetically appealing features (cf. Smith & Kirby, 2008). Alternatively, features that run against aesthetic preferences may acquire an advantage in learning and transmission, because they may elicit negative arousal and thus be more easily noticed and remembered (e.g., Citron, Weekes, & Ferstl, 2014; Kuperman, Estes, Brysbaert, & Waaarriner, 2014). In either case, aesthetic appeal (or lack thereof) could affect language learning, use, transmission, and change (e.g., sound changes or lexical borrowings; Rastall, 2008).

In this study, we investigate the fundamental premise of this proposal: that people’s aesthetic judgements of linguistic features or patterns consistently differ. To relate aesthetic judgements to language learning, we designed our study with respect to a widely investigated problem in language acquisition research: the speech segmentation problem (e.g., Saffran, Newport, & Aslin, 1996). This problem describes the challenge facing language learners to segment fluent speech into words. Several linguistic cues, including prosodic patterns in the speech stream, help learners solve this problem (e.g., Frost, Monaghan, & Tatsumi, 2017; Johnson & Jusczyk, 2001; Matzinger, Ritt, & Fitch, 2021; Ordin, Polyanskaya, Laka, & Nespor, 2017; Saffran et al., 1996; Tyler & Cutler, 2009). Therefore, we investigate listeners' aesthetic evaluation of different prosodic patterns, hypothesizing that the aesthetic appeal of prosodic patterns may correlate with the ease with which listeners can extract words from a continuous speech stream. If so, we could relate the aesthetic appeal of different prosodic patterns to speech segmentation, suggesting a role for "speech aesthetics" in language learning and change. However, any links identified between the aesthetic appeal of prosodic patterns and speech segmentation will represent correlations, but not causalities.

We chose prosodic patterns as targets for investigating the aesthetic appeal of linguistic features for two reasons. First, they play a crucial role in everyday speech segmentation and therefore support conclusions about general mechanisms of language learning and change. Second, prosody plays an important role in poetry and resembles musical patterns, whose aesthetic appeal has previously been investigated, making the link between aesthetics and prosody in everyday speech plausible (Hamilton, 2007; Obermeier et al., 2016).

Our study examined aesthetic evaluations of rhythmic patterns in trisyllabic words by native German speaking listeners, providing the necessary baseline for future research by explicitly testing if prosodic patterns differ in their aesthetic appeal at all (see discussion). We do not investigate links between aesthetic appeal and speech segmentation directly here, but rather refer to previous work on speech segmentation (e.g., Frost et al., 2017; Johnson & Jusczyk, 2001; Ordin et al., 2017; Tyler & Cutler, 2009), most notably to our own study (Matzinger et al., 2021), which closely resembled the present study in terms of stimuli, participants, and study design.

Our previous experiment used an artificial language learning setup to show that, when identifying trisyllabic words in a continuous speech stream, the performance of German speaking listeners improved when the final syllable of each word in the stream was lengthened (Matzinger et al., 2021). In contrast, their performance declined when the final syllable of each word was shortened. If aesthetic perception indeed plays a role in speech segmentation, words that have their final syllables lengthened or shortened should differ in terms of aesthetic appeal. More specifically, if speech segmentation is facilitated by aesthetically appealing prosodic patterns, words with lengthened final syllables should be rated highly for aesthetic appeal and words with shortened final syllables should be rated lower.

In addition, we found that native German speaking participants preferably extracted words with medially shortened syllables (Matzinger et al., 2021). Again, if speech segmentation is facilitated by a high aesthetic appeal, participants should perceive words with shortened medial syllables as aesthetically appealing.

However, durational changes might also be the result of language-specific stress patterns, in which stressed syllables are usually lengthened, and unstressed syllables shortened or reduced (Ordin et al., 2017; Tyler & Cutler, 2009). Due to the well-documented "mere exposure" effect (Sluckin, Hargreaves, & Colman, 1983; Zajonc, 1968), listeners may find durational variations that match the typical stress patterns of their native language more aesthetically appealing than other durational variations. Our participants were native speakers of German, and most German trisyllabic words are stressed on their word-medial syllable (Domahs, Plag, & Carroll, 2014; Ernestus & Neijt, 2008). Therefore, if participants prefer durational patterns that match the typical stress patterns of their native language, they should rank words with lengthened medial syllables as aesthetically appealing. However, alternative aesthetic theories suggest that moderately frequent, rare or novel items may be perceived as most aesthetically appealing (Berlyne, 1971; Martindale, Moore, & West, 1988; Temme, 1984), making the opposite prediction.

In sum, we investigated the aesthetic appeal of both word-*final* and word-*medial* prosodic patterns, because we can directly derive predictions from our previous speech segmentation experiment (Matzinger et al., 2021). In addition, we also tested the aesthetic appeal of words with word-*initial* durational modifications. However, neither our previous research, nor prosodic characteristics of German, implied any clear predictions for the aesthetic appeal of such words.

Finally, in addition to irregular durational patterns, we also examined the aesthetic appeal of isochronous words. Words with isochronous syllables may be regarded as aesthetically appealing because people have a general propensity for regular patterns (e.g., Poeppel & Assaneo, 2020; Ravignani & Madison, 2017), and isochrony has a facilitatory effect on auditory processing. On the other hand, people also perceive irregular patterns as aesthetically appealing (e.g., Westphal-Fitch & Fitch, 2013), and words with deviations from isochrony may be judged as more appealing than purely isochronous stimuli.

To summarize, our study investigated how listeners judged the aesthetic appeal of trisyllabic words where word-initial, medial, or final syllables were lengthened or shortened, compared to an isochronous baseline. We measured aesthetic appeal by collecting participants' ratings of these different prosodic patterns. They ranked each acoustic stimulus on its 'liking', its 'beauty' and its 'naturalness'. Although closely related, these concepts can represent different aspects of aesthetic appeal (Conway & Rehding, 2013). 'Liking' refers to purely sensual pleasure and is often regarded as an explicit evaluative judgement, whereas 'beauty' requires higher executive functions and includes more emotional involvement (Armstrong & Detweiler-Bedell, 2008; Brielmann & Pelli, 2017). ‘Naturalness’ ratings allowed us to test if judgements of likability and beauty are influenced by how natural participants find the stimuli.

Ratings of all three measures of aesthetic appeal might be influenced by the occurrence frequencies of the respective prosodic patterns in the participants’ native language, German (Sluckin et al., 1983). However, ‘naturalness’ is expected to be most strongly influenced by occurrence frequency, with more frequent prosodic patterns being judged as more natural (Zajonc, 1968). Therefore, naturalness ratings may help to determine if liking and beauty are influenced by the occurrence frequency of the respective prosodic patterns.

Methods

Experimental conditions and procedure

We examined whether participants perceived trisyllabic pseudo-words that either had isochronous syllables (isochrony condition) or had one of their syllables lengthened (lengthening conditions) or shortened (shortening conditions) as differing in aesthetic appeal. We conducted three otherwise identical experiments on different participants: in the first experiment, word-initial syllable durations were modified, the second experiment changed word-medial syllables, and the third modified word-final syllables. Durational condition (isochrony, lengthening, and shortening) was thus a within-subjects variable and modification position (word-initial, word-medial, and word-final) a between-subjects variable.

For each of 20 pseudo-word stimuli (Tab. 1), each participant rated its liking, beauty, and naturalness. Ratings were blocked: liking and beauty were counterbalanced in first and second blocks, and naturalness was always the last block. Thus, participants rated all stimuli on liking before rating the same stimuli on beauty (or vice versa), and finally provided naturalness ratings. All participants completed naturalness ratings last because naturalness ratings might be heavily biased by the stage of occurrence in the experiment (e.g., early in the experiment, artificial pseudo-words might be regarded as more unnatural than later in the experiment when participants have already been exposed to many similar pseudo-words) and because naturalness ratings served as a control category to test if participants' judgements of liking and beauty were influenced by their naturalness. Participants ranked each stimulus on a scale from 1 to 7, with 1 being the least and 7 being the most likable, beautiful, or natural.

For each experiment, each participant rated each word nine times, namely across three durational conditions (isochrony, lengthening, and shortening) and rating three manifestations of aesthetic appeal (liking, beauty, and naturalness). Modification positions (either word-initial, word-medial, or word-final position) varied across experiments, meaning that each participant only heard words modified in one position.

Participants and setting

Combining all 3 experiments, participants were 180 native or nativelike German speakers who spoke English as a second language. Each of the three experiments had 60 participants:

(a) word-initial modification: 13 male/43 female, mean age: 20.6 ± SD 1.98 years; 2 participants have not reported their gender and age; 2 participants were excluded because of technical issues during data collection.

(b) word-medial modification: 17 male/43 female, mean age: 20.5 ± SD 1.67 years

(c) word-final modification: 21 male/39 female, mean age: 21.4 ± SD 3.23 years

Participants were undergraduate psychology students at the University of Vienna. None reported auditory impairments. Participants were recruited via the online LABS system of the University of Vienna faculty of Psychology and received study credits for their participation.

Participants were tested in a laboratory setting free of background noise. Testing was administered via desktop computers; stimuli were presented binaurally over Sennheiser HD 300 PRO headphones at the same comfortable amplitude for all participants. Participants were seated approximately 60 cm from a 24'' monitor, and the experimental interface that was used to present the stimuli and collect ratings was created in OpenSesame (version 3.1.9; Mathôt, Schreij, & Theeuwes, 2012). Participants provided ratings by pressing number keys on a keyboard using their preferred hand.

After the experiment, participants provided demographic data and information about language background via a questionnaire. Testing was part of a longer testing session, and participation for the study described here lasted about 30 minutes per participant. Immediately after testing, there was a short debriefing and participants' questions about the study were answered. The study protocol was approved by the ethics board of the University of Vienna (reference number: #00489) and all participants gave written informed consent in accordance with the Declaration of Helsinki.

Stimuli

Creation of the artificial words

The artificial language used in these experiments consisted of 20 words (Tab. 1) consisting of three consonant-vowel (CV) syllables. Syllables were created from a pool of five German vowels (a, e, i, o, u) and thirteen consonants, namely six stops (b, d, g, p, t, k), three fricatives (f, v, s) and four sonorants (m, n, l, r). To avoid priming between artificial words, we ensured that each consonant-vowel syllable only occurred once in our set of words. Also, no sound occurred twice within a word and no consonant occurred in more than five words. The occurrence of vowels and consonants was balanced over the syllable positions so that each of the vowels appeared in each syllable position four times only, and no more than two words started with the same consonant. Moreover, we ensured that none of the created words was an existing word in German or English.

Some sounds (e.g., sonorants) may be more intrinsically pleasing than others (e.g., voiceless stops). Also, the lack of coarticulatory cues across syllables (see next section) may make some syllable transitions (e.g., vowel-nasal transitions) less natural than others. However, our balanced within-subjects design mitigates these potential confounds and our results did not reveal any systematic differences in the ratings across different words, suggesting that any potential intrinsic differences in the aesthetic appeal of individual syllables or transitions played at most a negligible role.

Table 1. Artificial words used in the study. Hyphens indicate syllable boundaries.

ba-pe-di	fa-ro-vu	ke-ta-fi	me-ko-ru	ri-fe-tu
bo-re-fu	fo-pu-ve	ku-te-so	mu-lo-se	si-go-va
da-ni-mo	ga-su-de	le-vi-po	ne-bu-pa	ti-nu-ge
do-mi-ka	gu-sa-ki	lu-bi-na	pi-ma-to	vo-la-gi

Stimulus creation

To create the sound stimuli, each CV syllable was recorded individually by a female native speaker of German. The acoustic parameters of the syllables were normalized in Praat (version 6.0.36; Boersma & Weenik, 2017) to a duration of 400 ms and a fundamental frequency (f_0) of 210 Hz, using Praat's pitch-synchronous overlap add (PSOLA) algorithm, and syllable amplitude was normalized by scaling syllable absolute peak amplitude to 99% of maximum. We checked that clipping was not an issue in any of our stimuli.

For the lengthening and shortening conditions, syllables were modified using PSOLA in Praat to lengths 150% or 50% of their original duration, yielding shortened syllables 200 ms long and lengthened syllables 600 ms long. The only stress correlate that varied across syllables was duration. Other stress correlates such as pitch or intensity were kept constant across syllables and therefore did not contribute to the perception of word stress in our stimuli (cf. Gordon & Roettger, 2017). Finally, the normalized and modified syllables were concatenated to form artificial words (Tab. 1) using custom code in Python 3.6.3.

Results

We preregistered a set of analyses (https://osf.io/6dfu5/), including ANOVAs and appropriate post-hoc tests, to compare the effects of durational condition on aesthetic appeal within the respective experiments, i.e. for each modification position separately. In addition to these pre-registered analyses, we used a Cumulative Link Mixed Model (CLMM, Christensen, 2018, 2019a; Christensen & Brockhoff, 2013) to combine the data from the three experiments. The CLMM reflects the categorical/ordinal nature of the rating scale data (Christensen &

Brockhoff, 2013), which is an advantage over standard linear mixed models. The CLMM, unlike the ANOVAs, also allows comparisons of the influence of durational modifications in different modification positions on the ratings of different prosodic patterns, and comparisons of the ratings of liking, beauty, and naturalness. Finally, the CLMM allows a control for non-independence of data points and random variation between participants by adding random effects, thus avoiding pseudo-replication and inflated type I error rates (Barr, Levy, Scheepers, & Tily, 2013).

We first provide a descriptive analysis of our results, then present the pre-registered ANOVA analyses, and finally discuss the explorative CLMM results. Statistics were run in R (version 3.6.0; R Development Core Team, 2018).

Descriptive statistics: mean liking, beauty, and naturalness ratings

Mean liking, beauty, and naturalness ratings (Fig. 1) revealed overall differences (indicated by non-overlapping confidence intervals) in aesthetic appeal between different durational modifications at different positions within trisyllabic words. In general, words with shortened syllables were rated as less aesthetically appealing than words with isochronous or lengthened syllables, and this effect manifested most prominently when shortening word-final syllables. However, these differences were small, with the mean ratings ranging from 2.83 (SD = 1.50; mean liking; word-finally shortened condition) to 3.83 (SD = 1.76; mean naturalness of isochronous syllables in the experiment on word-initial modifications) on a seven point rating scale (Fig. 1; Tab. 2). This is also reflected in rather low effect sizes (see ANOVA results below). Some participants showed more prominent differences between the three durational conditions than others, but the overall response patterns between participants were consistent, with shortened stimuli being consistently ranked lower than other stimuli (see Fig. S1 in the supplementary material).

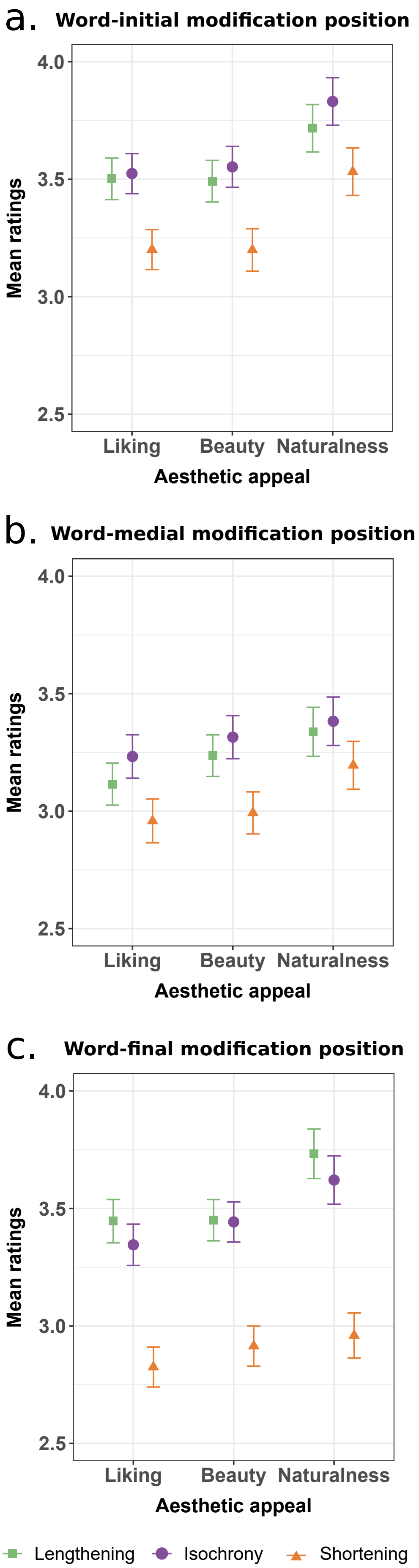

a.
Word-initial modification position
Mean ratings
4.0
3.5
3.0
2.5
Liking
Beauty
Naturalness
Aesthetic appeal
b.
Word-medial modification position
Mean ratings
4.0
3.5
3.0
2.5
Liking
Beauty
Naturalness
Aesthetic appeal
c.
Word-final modification position
Mean ratings
4.0
3.5
3.0
2.5
Liking
Beauty
Naturalness
Aesthetic appeal
Lengthening
Isochrony
Shortening

Figure 1. Mean liking, beauty, and naturalness ratings of words with isochronous, lengthened or shortened syllables in a) word-initial, b) word-medial and c) word-final position. Participants rated the stimuli's liking, beauty and naturalness on a scale from 1 (least likable, beautiful and natural) to 7 (most likable, beautiful and natural). Error bars denote 95% confidence intervals. Note that stimuli in the isochrony condition are identical across all three modification positions and were expected to be rated similarly. Since we used a between-subjects design for modification position, differences across the isochrony conditions may reflect the individual variation of participants taking part in the three experiments. Therefore, ratings on different durational modifications are best compared within a single modification position, and comparisons across modification positions interpreted cautiously.

Table. 2 Mean liking, beauty and naturalness ratings of words with isochronous, lengthened or shortened syllables in word-initial, word-medial and word-final position with standard deviations (SD), standard errors (SE) and 95 % confidence intervals (CI).

Modification position	**Aesthetic appeal**	**Durational condition**	**Mean**	**SD**	**SE**	**Lower CI**	**Upper CI**
Initial	**Liking**	Isochrony	3.52	1.48	0.04	3.44	3.61
		Lengthening	3.50	1.53	0.05	3.41	3.59
		Shortening	3.20	1.48	0.04	3.12	3.29
	Beauty	Isochrony	3.55	1.51	0.04	3.47	3.64
		Lengthening	3.49	1.54	0.05	3.40	3.58
		Shortening	3.20	1.57	0.05	3.11	3.29
	Naturalness	Isochrony	3.83	1.76	0.05	3.73	3.93
		Lengthening	3.72	1.75	0.05	3.62	3.82

		Shortening	3.53	1.75	0.05	3.43	3.63
Medial	**Liking**	Isochrony	3.23	1.63	0.05	3.14	3.32
		Lengthening	3.12	1.58	0.05	3.03	3.20
		Shortening	2.96	1.65	0.05	2.86	3.05
	Beauty	Isochrony	3.31	1.62	0.05	3.22	3.41
		Lengthening	3.24	1.57	0.05	3.15	3.32
		Shortening	2.99	1.57	0.05	2.90	3.08
	Naturalness	Isochrony	3.38	1.82	0.05	3.28	3.49
		Lengthening	3.34	1.84	0.05	3.23	3.44
		Shortening	3.19	1.80	0.05	3.09	3.30
Final	**Liking**	Isochrony	3.35	1.55	0.04	3.26	3.43
		Lengthening	3.45	1.63	0.05	3.35	3.54
		Shortening	2.83	1.50	0.04	2.74	2.91
	Beauty	Isochrony	3.44	1.50	0.04	3.36	3.53
		Lengthening	3.45	1.56	0.04	3.36	3.54
		Shortening	2.91	1.51	0.04	2.83	3.00
	Naturalness	Isochrony	3.62	1.82	0.05	3.52	3.72
		Lengthening	3.73	1.86	0.05	3.63	3.84
		Shortening	2.96	1.69	0.05	2.86	3.05

In general, mean ratings of liking and beauty were quite similar, whereas naturalness ratings were slightly higher (see non-overlapping confidence intervals in Fig. 1 and results of the CLMM below), indicating that participants’ evaluations of naturalness were more liberal than those for liking and beauty.

We investigated possible correlations between ratings for liking, beauty, and naturalness using Spearman's rank correlation coefficients (a non-parametric measure of the strength of the association between two ordinal variables; Spearman, 1910), using the cor.test function in R. This revealed a significant but moderate positive correlation (Cohen, 1992) between liking and beauty ratings ($r_s(10,678) = 0.50$, $p < 0.001$), liking and naturalness ratings ($r_s(10,678) = 0.46$, $p < 0.001$) and beauty and naturalness ratings ($r_s(10,678) = 0.46$, $p < 0.001$; Fig. 2). The degrees of freedom (10,678) are *n* – 2, with *n* being the number of ratings of each type of aesthetic appeal (178 participants * 20 words * 3 durations = 10,680 ratings).

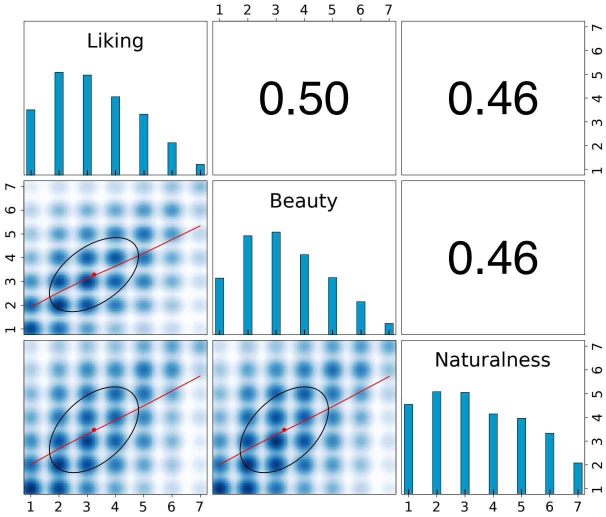


Figure 2: Correlations between liking, beauty and naturalness. Spearman's rank correlation coefficients are displayed in the top right panels, smoothed scatter plots in the bottom left panels, and histograms of the data distribution in the diagonal. Red points denote mean values and ellipses 95% confidence regions.

Confirmatory (pre-registered) analyses: ANOVAs

Analysis

For each of the three modification positions, we conducted three one-way repeated-measures ANOVAs (Field, Miles, & Field, 2012) to determine the effect of durational condition (independent variable) on liking, beauty and naturalness (dependent variables). We tested for sphericity (homogeneity of variances) using Mauchly's Test of Sphericity (Field, 1998; Field et al., 2012; Mauchly, 1940). Sphericity was violated in all of our ANOVAs (for W and p-values, see Tab. 3), so ANOVA results (degrees of freedom, F-values and p-values) were corrected using Greenhouse-Geisser correction (for ε values, see Tab. 3; Greenhouse & Geisser, 1959) to reduce the type I error rate (Field et al., 2012; Rouanet & Lépine, 1970). We used generalized eta square (Bakeman, 2005) to determine effect sizes.

Because all ANOVAs yielded statistically significant results, we conducted paired t-tests to make post-hoc comparisons between the respective groups, i.e. isochrony vs. lengthening, isochrony vs. shortening and lengthening vs. shortening (Field et al., 2012). We used a Bonferroni correction in all post-hoc tests, since the violation of sphericity can otherwise lead to biases in multiple comparisons (Boik, 1981; Keselman & Keselman, 1988; Maxwell, 1980). These Bonferroni corrections were not part of our pre-registered analyses, because the need for them only became evident after obtaining results for sphericity. All ANOVAs and post-hoc tests were conducted using the function *ezANOVA* of the R package "ez" (version 4.4-0; Lawrence, 2016) and a customized function based on the function *pairwise.t.test* from base R.

In all three experiments (all three modification positions), ANOVAs revealed a significant main effect of durational condition on liking, beauty, and naturalness (for degrees of freedom, F-values and p-values, see Tab. 3). Effect sizes for all effects were rather low (for generalized eta-squared measures η^2_g, see Tab. 3). Thus, modifying syllable duration either word-initially, word-medially or word-finally influenced the degree to which participants liked the presented words, and how beautiful and natural they found them, but only to a modest extent. Nonetheless, since this pattern was found in all three experiments, across different participants, it appears to be a robust and replicable effect.

Table 3. Results of the ANOVAs and Mauchly's Test for Sphericity, testing the effect of durational modification (independent variable) in three different modification positions on the liking, beauty and naturalness of words (dependent variable). df_{Num} indicates the degrees of freedom for the numerator, df_{Den} the degrees of freedom for the denominator, ε the Greenhouse-Geisser multiplier for degrees of freedom (degrees of freedom and p-values in the table incorporate the Greenhouse-Geisser correction). *F* indicates the F-values, p_A the p-values of the ANOVAs, and η^2_g the generalized eta-squared measures for effect size. *W* and p_M indicate W-values and p-values from Mauchly's Test for Sphericity.

Modification position	**Aesthetic appeal**	**ANOVA**							**Mauchly's Test for Sphericity**	
		Predictor	**df_{Num}**	**df_{Den}**	**ε**	**F**	**p_A**	**η^2_g**	**W**	**p_M**
Initial	**Liking**	Duration	1.45	82.47	0.72	16.33	<0.001	.03	0.62	<0.001
	Beauty	Duration	1.26	71.86	0.63	9.88	0.001	.03	0.41	<0.001
	Naturalness	Duration	1.36	77.69	0.68	7.05	0.005	.02	0.53	<0.001
Medial	**Liking**	Duration	1.45	85.83	0.73	5.81	0.009	.01	0.63	<0.001
	Beauty	Duration	1.54	90.61	0.77	12.71	<0.001	.02	0.70	<0.001

	Naturalness	Duration	1.45	85.39	0.72	3.93	0.036	.01	0.62	<0.001
Final	**Liking**	Duration	1.17	69.06	0.59	23.98	<0.001	.07	0.29	<0.001
	Beauty	Duration	1.17	69.06	0.59	23.98	<0.001	.07	0.44	<0.001
	Naturalness	Duration	1.21	71.38	0.60	26.98	<0.001	.10	0.35	<0.001

Results of the post-hoc paired t-tests (see Tab. 4) show that, regarding ‘liking’, words with shortened syllables at any position of the word were liked less than isochronous words, and that initially or finally shortened words were liked less than initially or finally lengthened words. Regarding ‘beauty’, words with shortened syllables at any position were regarded as less beautiful than isochronous words or words with lengthened syllables at the same position. With regard to ‘naturalness’, initially shortened words were rated as less natural than isochronous words, and finally shortened words as less natural than isochronous and finally lengthened words. We found no differences between isochronous stimuli and stimuli lengthened at any syllable position within the words with regard to liking, beauty or naturalness. Thus our results were driven by a general dislike of words with shortened syllables.

Table 4. Results from post-hoc paired t-tests, comparing whether isochronous words and words with lengthened or shortened syllables in three different modification positions differ in their likability, beauty and naturalness. *t*, *df* and *p* indicate t-values, degrees of freedom and p-values of the paired t-tests. All p-values are Bonferroni-corrected. Significant p-values are highlighted in bold.

Modification position	**Aesthetic appeal**	**Results from paired t-tests**			
		Comparison	**t**	**df**	**p**
Initial	**Liking**	isochrony vs. lengthening	-0.55	57	1.000
		isochrony vs. shortening	-4.89	57	**<0.001**
		lengthening vs. shortening	-3.91	57	**<0.001**
	Beauty	isochrony vs. lengthening	-1.34	57	0.553
		isochrony vs. shortening	-3.94	57	**<0.001**
		lengthening vs. shortening	-2.72	57	**0.026**
	Naturalness	isochrony vs. lengthening	-2.41	57	0.057
		isochrony vs. shortening	-3.45	57	**0.003**
		lengthening vs. shortening	-1.88	57	0.194
Medial	**Liking**	isochrony vs. lengthening	-2.09	59	0.122
		isochrony vs. shortening	-3.44	59	**0.003**

		lengthening vs. shortening	-1.56	59	0.371
	Beauty	isochrony vs. lengthening	-1.72	59	0.273
		isochrony vs. shortening	-4.59	59	**<0.001**
		lengthening vs. shortening	-3.07	59	**0.010**
	Naturalness	isochrony vs. lengthening	-1.04	59	0.914
		isochrony vs. shortening	-2.29	59	0.077
		lengthening vs. shortening	-1.84	59	0.214
Final	**Liking**	isochrony vs. lengthening	2.14	59	0.110
		isochrony vs. shortening	-5.23	59	**<0.001**
		lengthening vs. shortening	-4.96	59	**<0.001**
	Beauty	isochrony vs. lengthening	0.13	59	1.000
		isochrony vs. shortening	-6.02	59	**<0.001**

		lengthening vs. shortening	-4.50	59	**<0.001**
	Naturalness	isochrony vs. lengthening	2.05	59	0.130
		isochrony vs. shortening	-5.36	59	**<0.001**
		lengthening vs. shortening	-5.38	59	**<0.001**

Exploratory analysis: Cumulative Link Mixed Model

Analysis

To test whether the aesthetic perception of the artificial pseudo-words was influenced by *durational condition* and *modification position*, i.e. whether different durational modifications led to different ratings when they occurred in different positions, and whether this differed for *liking*, *beauty,* and *naturalness*, we applied a Cumulative Link Mixed Model (Christensen, 2018, 2019a; Christensen & Brockhoff, 2013), using Laplace approximation (Joe, 2008; Pinheiro & Bates, 1995).

In this model, durational condition, modification position and their interaction were included as fixed effects. The type of aesthetic rating (i.e. liking, beauty, and naturalness) was included as an additional fixed effect. An alternative model with a 3-way interaction between durational condition, modification position and aesthetic appeal had a higher AIC (Akaike Information Criterion) than our model, so we did not use it here. We also entered random intercepts of participant and pseudo-word into the model. To avoid inflated type I error rates, each model included a random slope (Barr et al., 2013) of durational condition within participant. The sample size for the model was 32,040 data points (178 individuals tested for 3 types of aesthetic appeal on 3 durational conditions with 20 pseudo-words each). We used the durational condition of isochrony, the word-medial modification position, and the aesthetic rating for naturalness as reference levels in the model.

The model was fitted in R (version 3.6.0; R Development Core Team, 2018), using the function *clmm* of the R-package "ordinal" (version 2019.12.10; Christensen, 2019b).

We used a likelihood ratio test to test the overall significance of the full model as compared to a null model comprising only the random effects (R function *anova*; Dobson, 2002). P-values for the effects of individual predictors are based on likelihood ratio tests that compare the full model with a reduced model lacking the fixed effects one at a time (R function *drop1*; Barr et al., 2013).

Results

Overall, the full model (Tab. 5) differed significantly from the null model, indicating an effect of durational condition, its potential interaction with modification position or of the type of aesthetic appeal on how participants rated the stimuli (likelihood ratio test: $\chi^2 = 212.0$, df = 10, $p < 0.001$). More specifically, we found that the durational condition influenced the ratings of stimuli (likelihood ratio test: $\chi^2 = 79.3$, df = 2, $p < 0.001$): in general, lengthening and shortening of syllables within a word had a negative effect on the ratings compared to isochrony (Tab. 5). Additionally, the interaction between durational condition and modification position had a significant effect on the ratings of stimuli (likelihood ratio test: χ^2

= 16.0, df = 4, p = 0.003), indicating that duration-based ratings were influenced by the syllable position at which they occurred. Word-finally lengthened syllables had an especially positive effect on the ratings (despite an overall negative effect of lengthening), whereas word-finally shortened syllables had an additional negative effect on the ratings (Tab. 5, Fig. 1a). Also, there was a significant effect of the type of aesthetic rating (likelihood ratio test: χ^2 = 112.1, df = 2, p < 0.001): participants provided lower ratings for liking and beauty relative to naturalness (Tab. 5). Overall, modification position did not affect the ratings (likelihood ratio test: χ^2 = 5.1, df = 2, p = 0.077), but participants in the experiment on word-initial modifications provided slightly higher ratings than participants in the experiment on medial modifications (Tab. 5). These results are consistent with the results of the pre-registered ANOVAs described above.

Table 5. Results of the Cumulative Link Mixed Model exploring the effects of duration, position, their interaction, and type of aesthetic appeal on the ratings of participants. The table reports estimated model coefficients (Estimate), standard errors (SE), z-values (z) and p-values (p) of the fixed effects, estimates, standard errors and z-values of the threshold coefficients, as well as variance, standard deviation (SD) and correlation coefficients (Corr) of the random effects.

Model coefficients	**Estimate**	**SE**	**z**	**p**	
appealLiking	-0.26	0.03	-10.28	**< 0.001**	
appealBeauty	-0.19	0.03	-7.56	**< 0.001**	
durationLonger	-0.11	0.05	-2.23	**0.026**	
durationShorter	-0.38	0.10	-3.68	**< 0.001**	
positionInitial	0.48	0.22	2.18	**0.029**	
positionFinal	0.25	0.22	1.13	0.257	

durationLonger:positionInitial	0.03	0.07	0.41	0.684	
durationShorter:positionInitial	-0.07	0.15	-0.44	0.659	
durationLonger:positionFinal	0.20	0.07	2.92	**0.004**	
durationShorter:positionFinal	-0.44	0.15	-3.04	**0.002**	
Threshold coefficients	**Estimate**	**SE**	**z**		
1|2	-2.40	0.20	-12.23		
2|3	-0.82	0.20	-4.18		
3|4	0.34	0.20	1.73		
4|5	1.30	0.20	6.64		
5|6	2.39	0.20	12.18		
6|7	3.86	0.20	19.49		
Random effects	**Term**	**Variance**	**SD**	**Corr**	
participant	**Intercept**	1.37	1.17		
	durationLonger	0.04	0.19	0.06	
	durationShorter	0.52	0.72	-0.25	-0.98
word	**Intercept**	0.29	0.54		

Discussion

Our experiments yielded clear and consistent effects of modifications of syllable duration in nonsense words on different measures of their aesthetic appeal. These results are consistent with previously published research on speech segmentation, and thus overall with our initial hypothesis that there is a relationship between the aesthetic appeal of words and the ease with which they are processed.

Specifically, pre-registered confirmatory analyses (ANOVAs and post-hoc tests) revealed that words with one syllable shortened were generally perceived as less aesthetically appealing than isochronous words and words with one syllable lengthened, independently of the syllable position at which the modification occurred. More rigorous exploratory analyses (CLMM) confirmed this negative effect of shortening, and indicated that it was especially prominent when shortening occurred word-finally. Additionally, words with a syllable lengthened had a lower aesthetic appeal than isochronous words, but only when lengthening occurred word-initially or word-medially, and not when it occurred word-finally (significant interaction effect of durational modification and modification position). Overall, these findings are consistent with a general human preference for regular and isochronous patterns in words (e.g. Ravignani & Madison, 2017).

The finding that words with word-finally shortened syllables had a particularly low aesthetic appeal (see Fig. 1 and CLMM results) is in line with our previous speech segmentation experiment (Matzinger et al., 2021), in which final-syllable shortening hindered the segmentation of words from a continuous speech stream. Thus, low aesthetic appeal of prosodic patterns correlates with a disadvantage of these patterns for speech segmentation, potentially influencing language learning in a wider sense. The causalities behind these correlations remain unclear: they could indicate either that listeners do not use aesthetically unpleasant prosodic patterns for speech segmentation, or that patterns that are not used for speech segmentation for other reasons are not perceived as aesthetically appealing, or both. Additional underlying factors, such as acoustic salience, speech rate, context, occurrence frequencies, neural oscillations, ease of processing or memory may also influence both appeal and segmentation (Dilley & Pitt, 2010; Forster, Leder, & Ansorge, 2013; Morrill et al., 2015; Obermeier et al., 2016; Palmer & Mattys, 2016; Poeppel & Assaneo, 2020; Reber et al., 2004). For example, shortened syllables could indicate faster speech, which may be less preferred (but this would predict that lengthening should consistently increase preferences, which was not the case). Notably, there may also be factors that only influence speech segmentation but have no effect on aesthetic appeal, or vice versa. Pinning down such additional influencing factors is a topic for future research.

Finally lengthened syllables boosted speech segmentation in our previous study (Matzinger et al., 2021), but these words did not differ in their aesthetic appeal from isochronous words (see Fig. 1 and Tab. 4). This indicates that the positive effect of word-final lengthening for speech segmentation may not be directly related to its aesthetic appeal, but rather to other underlying factors (see above). Alternatively, participants may have factored out the effects of final lengthening because they perceived the words as highly natural. Still, the finally-lengthened patterns that facilitated speech segmentation were more aesthetically pleasing than the finally-shortened patterns that hindered it, which again suggests a link between aesthetic appeal and speech segmentation success. Overall, our results for finally lengthened and shortened words suggest that prosodic patterns which support speech segmentation may not necessarily need to be aesthetically pleasing, but that unaesthetic patterns hinder speech segmentation, and potentially language learning in a more general sense.

In our previous speech segmentation experiment (Matzinger et al., 2021), shortened syllables were preferably segmented as occurring word-medially. This led to the prediction for this study that, if speech segmentation correlates with liking, shortened medial syllables should be ranked as most aesthetically appealing. This prediction, however, was not borne out because – like in word-initial and word-final position – shortened syllables were perceived as less likable and beautiful than lengthened and isochronous syllables (Fig. 1; Tab. 4). Still, the negative effect of medial shortening on likability and beauty was smaller than the negative effect for initial and final shortening, and there was no negative effect of medial shortening on naturalness (Fig. 1). This may reflect that initial and final syllables are particularly important for speech processing and segmentation, while word-medial syllables are less salient (Hall, Hume, Jaeger, & Wedel, 2018; Tyler & Cutler, 2009; Wedel, Ussishkin, & King, 2019). Although the evidence for this is indirect, this finding may also reflect a link between aesthetic appeal and speech segmentation performance.

Interestingly, naturalness ratings of our prosodic patterns did not correlate with frequencies of stress patterns in the participants' native language. Since most German trisyllables are stressed on the word-medial syllable, and stress usually correlates with lengthening (Domahs et al., 2014; Ernestus & Neijt, 2008; Ordin et al., 2017), we expected words with lengthened word-medial syllables to be most natural, and words with shortened word-medial syllables least natural. Instead, we found that naturalness ratings of medially lengthened and medially shortened words was comparable. This may be because although the majority of German trisyllables are stressed word-medially, German also has trisyllabic words that do not carry word-medial stress. Listeners may not just rate the majority pattern as most natural but may consider the distributions of different patterns when evaluating their naturalness. Interestingly, there was a large difference in naturalness ratings of words with finally lengthened and finally shortened syllables: finally lengthened words were rated as much more natural. This finding may be explained by the occurrence frequencies of prosodic patterns in natural languages. Cross-linguistically, final syllables are lengthened for multiple reasons, for example to indicate boundaries (Fletcher, 2010). This boundary-related lengthening mostly happens phrase-finally, and not word-finally, but since in our design, each word essentially was a phrase, participants may have transferred the high naturalness of phrase-final lengthening to our stimuli. This also suggests that listeners may be primed more by cross-linguistic than language-specific prosodic patterns when evaluating their aesthetic appeal.

We did not evaluate other prosodic cues, such as variations in f0 or intensity, in our stimuli, which allows us to clearly attribute the differences in aesthetic appeal to our durational modifications. However, keeping all other prosodic cues fixed leads our stimuli to sound slightly unnatural. In natural languages, stress has multiple correlates, and participants in our experiments may not have interpreted syllables that had a longer duration, but not a higher intensity or a higher f0, as being stressed (cf. Gordon & Roettger, 2017). Also, in natural languages, pitch and intensity typically decrease phrase-finally, which was not true of our stimuli (Vaissière, 1983). Future experiments evaluating the aesthetic appeal of prosodic patterns could combine different voice modulatory cues, which would resemble natural speech more closely and provide a more nuanced, but complex, picture. Also, to make the task more natural from the outset, target words could be embedded in a larger context (e.g., in a framing utterance) in future experiments.

Liking, beauty and naturalness ratings were positively correlated, indicating that our participants perceived these three concepts to be highly related. However, naturalness ratings were consistently slightly higher than liking and beauty ratings, suggesting that they were provided more forgivingly than liking and beauty ratings. Participants possibly did not find our stimuli particularly appealing in general (e.g. for reasons discussed above), but this dislike may be less reflected in the naturalness ratings because naturalness may be influenced less by explicit aesthetic judgements than liking and beauty. Alternatively, in our study design, naturalness ratings were always provided last, and participants may have provided higher ratings after a longer exposure to the stimuli (Sluckin 1983). Future studies could thus potentially use a single measure of aesthetic appeal, rather than the three used here.

Disentangling the concepts of liking, beauty and naturalness was not the main focus of our study. Instead, we used the three concepts in an exploratory fashion, to gain initial insights into possible correlations or differences between them. Thus, the correlations between the three different dimensions of aesthetic appeal may reflect that the boundaries between them are too fuzzy for participants to differentiate between them without prior instruction. We do not consider this as problematic because we aimed to capture subjective interpretations of these three concepts. Future studies specifically aiming at separating different dimensions of aesthetic appeal (as often done in the visual domain; Leder et al., 2013; Lyssenko, Redies, & Hayn-Leichsenring, 2016; Sidhu, McDougall, Jalava, & Bodner, 2018) should ensure that participants agree about the three qualitative dimensions of aesthetic appeal, for example by defining the concepts before the experiment or by integrating a training phase and including an inter-rater reliability analysis.

The current study provides a proof of concept that durational modifications influence aesthetic ratings, and that the effects are consistent with speech segmentation results, but our experiments do not definitively show that aesthetic judgments in turn influence speech processing (or by extension, language change). It would be valuable to investigate in future studies whether the subjective dimensions of aesthetic appeal correlate with more objective measures such as results from very simple syllable or word recognition tasks (Ferrand et al., 2018; Goldinger, 1996), which are cognitively less demanding than the speech segmentation task of Matzinger et al. (2021). For example, an experiment could test if participants better remember or recognize words that they rated as aesthetically appealing. If correlations between aesthetic appeal and simple learning tasks exist, these simple learning tasks might provide more immediate proxies for evaluating the effects of aesthetic appeal on other aspects of language learning and hence historical language change.

Conclusion

Our results demonstrate that different prosodic patterns differ in their likability, beauty, and naturalness. This finding provides an important baseline for further investigating the potential relationship between aesthetic appeal, prosodic patterns, and cognitive factors such as learnability, ease of processing or memory. Together with previous work on the effectiveness of different prosodic cues for speech segmentation, our findings on aesthetic appeal render it plausible that such a relationship exists. Our study thus establishes a crucial prerequisite for the idea that aesthetic perception can act as a potential biasing force in language learning and language change, opening up new avenues for research. Deeper insights into the role of aesthetic perception of linguistic features as a potential bias in language learning and language change will require further studies that test this relationship directly (e.g. using iterated learning experiments).

Data availability statement

All data and R files used for the analysis are available in the Open Science Framework repository and can be accessed at https://osf.io/6dfu5/.

Acknowledgements

We would like to thank Maximilian Douda for helping with the data collection.

Funding

This work was supported by the Austrian Science Fund (FWF) DK Grant Cognition & Communication [grant number: #W1262-B29] to WTF.

References

Armstrong, T., & Detweiler-Bedell, B. (2008). Beauty as an Emotion: The Exhilarating Prospect of Mastering a Challenging World. *Review of General Psychology*, *12*(4), 305–329. https://doi.org/10.1037/a0012558

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, *37*(3), 379–384. https://doi.org/10.3758/BF03192707

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Berlyne, D. E. (1971). *Aesthetics and psychobiology*. New York: Appleton Century Crofts.

Boersma, P., & Weenik, D. (2017). Praat: doing phonetics by computer. Retrieved from http://www.praat.org/

Boik, R. J. (1981). A priori tests in repeated measures designs: Effects of nonsphericity. *Psychometrika*, *46*(3), 241–255. https://doi.org/10.1007/BF02293733

Brielmann, A. A., & Pelli, D. G. (2017). Beauty Requires Thought. *Current Biology*, *27*(10), 1506-1513.e3. https://doi.org/10.1016/j.cub.2017.04.018

Christensen, R. H. B. (2018). Cumulative link models for ordinal regression with the R package ordinal.

Christensen, R. H. B. (2019a). A Tutorial on fitting Cumulative Link Models with the ordinal Package. Retrieved from https://cran.r-project.org/web/packages/ordinal/vignettes/clmm2_tutorial.pdf

Christensen, R. H. B. (2019b). ordinal - Regression Models for Ordinal Data. Retrieved from https://cran.r-project.org/package=ordinal

Christensen, R. H. B., & Brockhoff, P. (2013). Analysis of sensory ratings data with cumulative link models. *Journal de La Société Française de Statistique & Revue de Statistique Appliquée*, *154*(3), 58–79.

Citron, F. M. M., Weekes, B. S., & Ferstl, E. C. (2014). Arousal and emotional valence interact in written word recognition. *Language, Cognition and Neuroscience*, *29*(10), 1257–1267. https://doi.org/10.1080/23273798.2014.897734

Cohen, J. (1992). A Power Primer. *Psychological Bulletin*, *112*(1), 155–159.

Conway, B. R., & Rehding, A. (2013). Neuroaesthetics and the Trouble with Beauty. *PLoS Biology*, *11*(3), 1–5. https://doi.org/10.1371/journal.pbio.1001504

Dilley, L. C., & Pitt, M. A. (2010). Altering context speech rate can cause words to appear or disappear. *Psychological Science*, *21*(11), 1664–1670. https://doi.org/10.1177/0956797610384743

Dobson, A. J. (2002). *An introduction to generalized linear models*. Boca Raton: Chapman & Hall.

Domahs, U., Plag, I., & Carroll, R. (2014). Word stress assignment in German, English and Dutch: Quantity-sensitivity and extrametricality revisited. *The Journal of Comparative Germanic Linguistics*, *17*(1), 59–96. https://doi.org/10.1007/s10828-014-9063-9

Ernestus, M., & Neijt, A. (2008). Word length and the location of primary word stress in Dutch, German, and English. *Linguistics*, *46*(3), 507–540. https://doi.org/10.1515/LING.2008.017

Ferrand, L., Méot, A., Spinelli, E., New, B., Pallier, C., Bonin, P., … Grainger, J. (2018).

MEGALEX: A megastudy of visual and auditory word recognition. *Behavior Research Methods*, *50*(3), 1285–1307. https://doi.org/10.3758/s13428-017-0943-1

Field, A. (1998). A Bluffer ' s Guide to ... Sphericity. *British Psychological Society - Mathematics, Statistics, and Computing Newsletter*, *6*(1), 13–24.

Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics Using R*. *International Statistical Review*. Los Angeles: SAGE. https://doi.org/10.1111/insr.12011_21

Fletcher, J. (2010). The Prosody of Speech : Timing and Rhythm. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The handbook of phonetic sciences* (2nd ed., pp. 523–602). Hoboken: Wiley-Blackwell.

Foolen, A., Lüdtke, U. M., Racine, T. P., & Zlatev, J. (Eds.). (2012). *Moving ourselves, moving others*. John Benjamins Publishing Company.

Forster, M. (2020). Processing fluency. In M. Nadal & O. Vartanian (Eds.), *The Oxford Handbook of Empirical Aesthetics*. Oxford: Oxford University Press.

Forster, M., Leder, H., & Ansorge, U. (2013). It felt fluent, and i liked it: Subjective feeling of fluency rather than objective fluency determines liking. *Emotion*, *13*(2), 280–289. https://doi.org/10.1037/a0030115

Frost, R. L. A., Monaghan, P., & Tatsumi, T. (2017). Domain-general mechanisms for speech segmentation: The role of duration information in language learning. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(3), 466–476. https://doi.org/10.1037/xhp0000325

Goldinger, S. D. (1996). Auditory lexical decision. *Language and Cognitive Processes*, *11*(6), 559–568. https://doi.org/10.1080/016909696386944

Gordon, M., & Roettger, T. (2017). Acoustic correlates of word stress: A cross-linguistic survey. *Linguistics Vanguard*, *3*(1), 1–11. https://doi.org/10.1515/lingvan-2017-0007

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, *24*(2), 95–112.

Hall, K. C., Hume, E., Jaeger, T. F., & Wedel, A. (2018). The role of predictability in shaping phonological patterns. *Linguistics Vanguard*, *20170027*, 1–15. https://doi.org/10.1515/lingvan-2017-0027

Hamilton, A. (2007). *Aesthetics and music*. London: Continuum International Publishing Group.

Joe, H. (2008). Accuracy of Laplace approximation for discrete response mixed models. *Computational Statistics and Data Analysis*, *52*(12), 5066–5074. https://doi.org/10.1016/j.csda.2008.05.002

Johnson, E. K., & Jusczyk, P. W. (2001). Word Segmentation by 8-Month-Olds : When Speech Cues Count More Than Statistics. *Journal of Memory and Language*, *44*, 548–567. https://doi.org/10.1006/jmla.2000.2755

Keller, R. (1994). *On language change - the invisible hand in language*. New York: Routledge. https://doi.org/10.4324/9780203993286

Keselman, H. J., & Keselman, J. C. (1988). Comparing repeated measures means in factorial designs. *Psychophysiology*, *25*(5), 612–618.

Kousta, S. T., Vinson, D. P., & Vigliocco, G. (2009). Emotion words, regardless of polarity, have a processing advantage over neutral words. *Cognition*, *112*(3), 473–481. https://doi.org/10.1016/j.cognition.2009.06.007

Kuperman, V., Estes, Z., Brysbaert, M., & Warriner, A. B. (2014). Emotion and language: valence and arousal affect word recognition. *Journal of Experimental Psychology. General*, *143*(3), 1065–1081. https://doi.org/10.1037/a0035669.Emotion

Laham, S. M., Koval, P., & Alter, A. L. (2012). The name-pronunciation effect: Why people like Mr. Smith more than Mr. Colquhoun. *Journal of Experimental Social Psychology*, *48*(3), 752–756. https://doi.org/10.1016/j.jesp.2011.12.002

Lawrence, M. A. (2016). ez: Easy Analysis and Visualization of Factorial Experiments. Retrieved from https://cran.r-project.org/package=ez

Leder, H., Ring, A., & Dressler, S. G. (2013). See me, feel me! Aesthetic evaluations of art portraits. *Psychology of Aesthetics, Creativity, and the Arts*, *7*(4), 358–369. https://doi.org/10.1037/a0033311

Lyssenko, N., Redies, C., & Hayn-Leichsenring, G. U. (2016). Evaluating abstract art: Relation between term usage, subjective ratings, image properties and personality traits. *Frontiers in Psychology*, *7*(JUN), 1–9. https://doi.org/10.3389/fpsyg.2016.00973

Martindale, C., Moore, K., & West, A. (1988). Relationship of Preference Judgments to Typicality, Novelty, and Mere Exposure. *Empirical Studies of the Arts*, *6*(1), 79–96. https://doi.org/10.2190/mcaj-0gqt-djtl-lnqd

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7

Matzinger, T., Ritt, N., & Fitch, W. T. (2021). The Influence of Different Prosodic Cues on Word Segmentation. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.622042

Mauchly, J. W. (1940). Significance test for sphericity of a normal n-variate distribution. *The Annals of Mathematical Statistics*, *11*, 204–209. Retrieved from http://projecteuclid.org/euclid.aop/1176996548

Maxwell, S. E. (1980). Pairwise Multiple Comparisons in Repeated Measures Designs. *Journal of Educational Statistics*, *5*(3), 269–287. https://doi.org/10.3102/10769986005003269

Morrill, T. H., McAuley, J. D., Dilley, L. C., Zdziarska, P. A., Jones, K. B., & Sanders, L. D. (2015). Distal prosody affects learning of novel words in an artificial language. *Psychonomic Bulletin and Review*, *22*(3), 815–823. https://doi.org/10.3758/s13423-014-0733-z

Nadal, M., & Vartanian, O. (Eds.). (2019). *The Oxford Handbook of Empirical Aesthetics*.

Oxford: Oxford University Press.

Obermeier, C., Kotz, S. A., Jessen, S., Raettig, T., von Koppenfels, M., & Menninghaus, W. (2016). Aesthetic appreciation of poetry correlates with ease of processing in event-related potentials. *Cognitive, Affective and Behavioral Neuroscience*, *16*(2), 362–373. https://doi.org/10.3758/s13415-015-0396-x

Ordin, M., Polyanskaya, L., Laka, I., & Nespor, M. (2017). Cross-linguistic differences in the use of durational cues for the segmentation of a novel language. *Memory & Cognition*, *45*, 863–876. https://doi.org/10.3758/s13421-017-0700-9

Palmer, S. D., & Mattys, S. L. (2016). Speech segmentation by statistical learning is supported by domain-general processes within working memory. *Quarterly Journal of Experimental Psychology*, *69*(12), 2390–2401. https://doi.org/10.1080/17470218.2015.1112825

Paulmann, S., Bleichner, M., & Kotz, S. A. (2013). Valence, arousal, and task effects in emotional prosody processing. *Frontiers in Psychology*, *4*(June), 1–10. https://doi.org/10.3389/fpsyg.2013.00345

Pinheiro, J. C., & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, *4*(1), 12–35. https://doi.org/10.1080/10618600.1995.10474663

Poeppel, D., & Assaneo, M. F. (2020). Speech rhythms and their neural foundations. *Nature Reviews Neuroscience*, *21*(6), 322–334. https://doi.org/10.1038/s41583-020-0304-4

R Development Core Team. (2018). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org/

Rastall, P. (2008). Aesthetic responses and the “cloudiness” of language: Is there an aesthetic function of language? *La Linguistique*, *44*, 103–132. https://doi.org/10.3917/ling.441.0103

Ravignani, A., & Madison, G. (2017). The paradox of isochrony in the evolution of human rhythm. *Frontiers in Psychology*, *8*, 1–13. https://doi.org/10.3389/fpsyg.2017.01820

Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality and Social Psychology Review*, *8*(4), 364–382. https://doi.org/10.1207/s15327957pspr0804_3

Rouanet, H., & Lépine, D. (1970). Comparison between treatments in a repeated-measures design: ANOVA and multivariate methods. *The British Journal of Mathematical and Statistical Psychology*, *23*(2), 147–163.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word Segmentation : The Role of Distributional Cues. *Journal of Memory and Language*, *35*, 606–621.

Shibles, W. (1995). *Emotion in aesthetics*. Dordrecht: Springer Science + Business Media.

Sidhu, D. M., McDougall, K. H., Jalava, S. T., & Bodner, G. E. (2018). Prediction of beauty and liking ratings for abstract and representational paintings using subjective and objective measures. *PLoS ONE*, *13*(7), 1–15. https://doi.org/10.1371/journal.pone.0200431

Sluckin, W., Hargreaves, D. J., & Colman, A. M. (1983). Novelty and human aesthetic preferences. In J. Archer & L. I. A. Birke (Eds.), *Exploration in animals and humans* (pp. 245–269). Wokingham: Van Nostrand Reinhold. Retrieved from papers2://publication/uuid/1BB53451-CC54-4C2A-8F4A-52A3A30A7B88

Smith, K., & Kirby, S. (2008). Cultural evolution: implications for understanding the human language faculty and its evolution. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *363*(1509), 3591–3603. https://doi.org/10.1098/rstb.2008.0145

Spearman, C. (1910). Correlation Calculated From Faulty Data. *British Journal of Psychology, 1904-1920*, *3*(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Temme, J. E. (1984). Effects of Mere Exposure, Cognitive Set and Task Expectations on Aesthetic Appreciation. In W. R. Crozier & C. A. J (Eds.), *Cognitive Processes in the Perception of Art* (pp. 389–410). North-Holland: Elsevier. https://doi.org/10.1016/S0166-4115(08)62360-2

Tyler, M. D., & Cutler, A. (2009). Cross-language differences in cue use for speech segmentation. *Journal of the Acoustical Society of America*, *126*(1), 367–376. https://doi.org/10.1121/1.3129127

Vaissière, J. (1983). Language-Independent Prosodic Features. In A. Cutler & D. R. Ladd (Eds.), *Springer Series in Language and Communication 14: Prosody: Models and Measurements* (pp. 53–66). Hamburg: Springer.

Verheyen, L. (2015). The aesthetic experience of the literary artwork. A matter of form and content? *Aesthetic Investigations*, *1*(1), 23–32. Retrieved from http://www.aestheticinvestigations.eu/index.php/journal/article/view/37

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

Wedel, A., Ussishkin, A., & King, A. (2019). Incremental word processing influences the evolution of phonotactic patterns. *Folia Linguistica*, *40*(1), 231–248. https://doi.org/10.1515/flih-2019-0011

Westphal-Fitch, G., & Fitch, W. T. (2013). Spatial Analysis of “Crazy Quilts”, a Class of Potentially Random Aesthetic Artefacts. *PLoS ONE*, *8*(9). https://doi.org/10.1371/journal.pone.0074055

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, *9*(2), 1–27. Retrieved from http://psycnet.apa.org/journals/psp/9/2p2/1/

Aesthetic appeal of prosodic patterns influences their segmentation from a speech stream

Theresa Matzinger, Eva Specker, Nikolaus Ritt, W. Tecumseh Fitch

Supplementary material

Between-subject variability of the rankings

Some participants showed more prominent differences between the three durational conditions than others, but the overall response patterns between participants were consistent, with shortened stimuli being consistently ranked lower than other stimuli (Fig. S1). Overall, participants used the whole rating scale, with some participants mainly using the lower end of the rating scale, and others mainly used the middle or the higher end of the rating scale (Fig. S1). In general, the highest ratings occurred only rarely and lower ratings were featured more prominently. This is reflected in the mean values for naturalness being centered around the middle of the rating scale (3.5) and those for liking and beauty slightly below (Fig. 1).

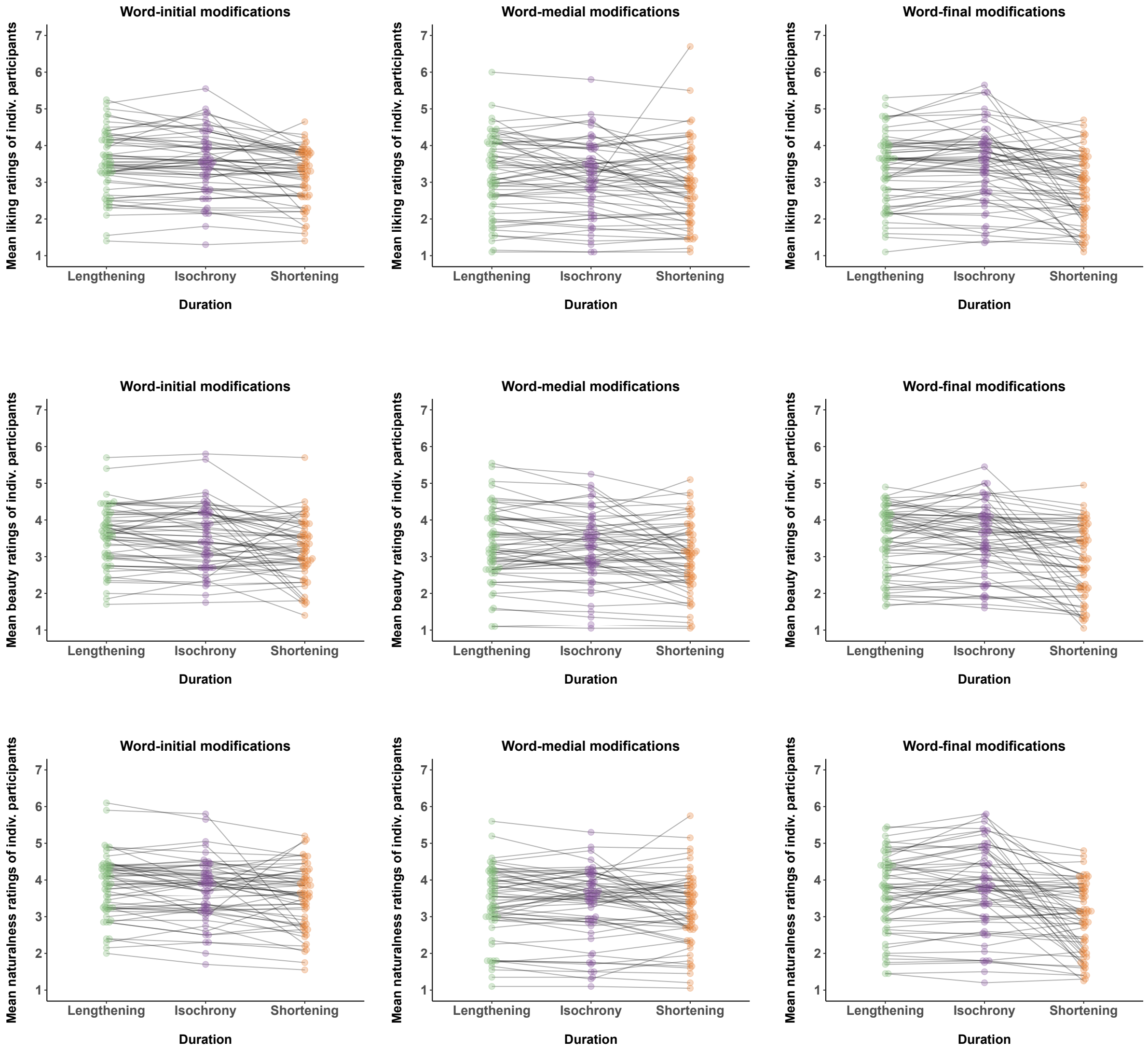


Figure S1. Mean liking (first row), beauty (second row) and naturalness (third row) ratings of individual participants of words with initially (first column), medially (second column) or finally (third column) modified syllables. Gray lines connect the ratings of individual participants of words with isochronous, lengthened and shortened syllables.

CHAPTER 4

Non-native speaker pause patterns closely correspond to those of native speakers at different speech rates

This chapter is published in *PLoS ONE*.

Citation:

Matzinger T, Ritt N & Fitch WT. 2020. Non-native speaker pause patterns closely correspond to those of native speakers at different speech rates. *PLoS ONE* 15(4), e0230710. https://doi.org/10.1371/journal.pone.0230710

PLOS ONE

RESEARCH ARTICLE

Non-native speaker pause patterns closely correspond to those of native speakers at different speech rates

Theresa Matzinger[1,2]*, **Nikolaus Ritt**[1], **W. Tecumseh Fitch**[2]*

1 Department of English, University of Vienna, Vienna, Austria, **2** Department of Cognitive Biology, University of Vienna, Vienna, Austria

* theresa.matzinger@univie.ac.at (TM); tecumseh.fitch@univie.ac.at (WTF)

Abstract

When speaking a foreign language, non-native speakers can typically be readily identified by their accents. But which aspects of the speech signal determine such accents? Speech pauses occur in all languages but may nonetheless vary in different languages with regard to their duration, number or positions in the speech stream, and therefore are one potential contributor to foreign speech production. The aim of this study was therefore to investigate whether non-native speakers pause 'with a foreign accent'. We recorded native English speakers and non-native speakers of German or Serbo-Croatian with excellent English reading out an English text at three different speech rates, and analyzed their vocal output in terms of number, duration and location of pauses. Overall, all non-native speakers were identified by native raters as having non-native accents, but native and non-native speakers made pauses that were similarly long, and had similar ratios of pause time compared to total speaking time. Furthermore, all speakers changed their pausing behavior similarly at different speech rates. The only clear difference between native and non-native speakers was that the latter made more pauses than the native speakers. Thus, overall, pause patterns contributed little to the acoustic characteristics of speakers' non-native accents, when reading aloud. Non-native pause patterns might be acquired more easily than other aspects of pronunciation because pauses are perceptually salient and producing pauses is easy. Alternatively, general cognitive processing mechanisms such as attention, planning or memory may constrain pausing behavior, allowing speakers to transfer their native pause patterns to a second language without significant deviation. We conclude that pauses make a relatively minor contribution to the acoustic characteristics of non-native accents.

OPEN ACCESS

Citation: Matzinger T, Ritt N, Fitch WT (2020) Non-native speaker pause patterns closely correspond to those of native speakers at different speech rates. PLoS ONE 15(4): e0230710. https://doi.org/10.1371/journal.pone.0230710

Editor: Koji Miwa, Nagoya University, JAPAN

Received: July 19, 2019

Accepted: March 6, 2020

Published: April 3, 2020

Data Availability Statement: All relevant data as well as the Praat scripts and R scripts used for data processing and analysis can be accessed at the Open Science Framework: DOI 10.17605/OSF.IO/JC7EF.

Funding: Preparation of this paper was supported by Austrian Science Fund (FWF; https://www.fwf.ac.at) DK Grant "Cognition & Communication" (#W1262-B29). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

When speaking a foreign language, most non-native speakers can be readily identified by their accents, which are often distinctive depending on native language. To understand how different accents arise, and also to help second language learners to eliminate them to improve intelligibility, it is essential to know which aspects of spoken language contribute to non-native speech production.

When studying foreign accents, it is crucial to distinguish between accents in production, i.e. atypical measurable acoustic characteristics of the speech signal, and accents in perception, i.e. listeners' identification of accents as non-native [1–3]. Although these factors are probably related, they are not identical, and the focus of the current study is solely on accents in production.

A crucial factor generating distinct foreign accents is that linguistic phenomena vary between languages, and speakers transfer language-specific features from one language to another (e.g. [4]). These language-typical features can range from pronunciation of individual phonemes (e.g. [5,6]) to suprasegmental prosodic features like intonation patterns (e.g. [7,8]). For example, native speakers of Japanese have difficulty distinguishing the phonemes /l/ and /r/ in foreign languages because Japanese has only one similar phoneme (e.g.[9]). Native German learners of English often substitute /d/ or /s/ for /ð/ or /θ/ (as in *this* or *thing*) because the latter two English sounds are not part of the German phoneme inventory [10]. German speakers also typically use a different pitch range than English, making native German speakers of English sound "bored" to English native speakers, whereas native English speakers often sound "over-excited" to German native speakers [11–13]. Such examples could be readily multiplied.

Certain linguistic features seem to contribute less to foreign speech production than others. These features include linguistic phenomena claimed to be language-universal [14] that are transferred to (or acquired easily in) a second language (L2). Such language-universal phenomena are rare, and have been suggested to mainly concern pragmatic, rather than phonological or grammatical, features [15–18].

The current study investigates whether L1 and L2 speakers differ in their pausing behavior when reading aloud, and hence whether different pausing patterns contribute to a 'foreign accent' in speech production. Pauses are particularly interesting candidates to investigate non-native accents for several reasons. Due to the natural need to breathe, pauses occur in all of the world's languages. Still, not all pauses are determined by physiological needs, but have functions closely related to cognition: pauses help structure speech, determine speech tempo and rhythm, plan upcoming utterances, can add rhetorical emphasis, or structure turn-taking [19–23]. Such roles, reflected for example by the durations of pauses or by the positions of pauses in a stretch of speech, may be completely physiologically determined or may be realized similarly in different languages (e.g. [24–26,27] see S1 Table) and thus either transfer without change, or be easy to acquire, and thus not contribute to non-native speech production.

On the other hand, the different functions of pauses potentially make them subject to cross-linguistic variation (see S2 Table). For example, there is evidence that English speakers pause more frequently than French [28,29] or Turkish speakers [30], but less frequently than Spanish speakers [31]. Also, English speakers' pauses are shorter than French speakers' [28,32,but also 33] and Russian speakers' [34], but longer than Italian speakers' pauses [33]. Such cross-linguistic differences in pause patterns, if carried over to a foreign language, might contribute to non-native accents in speech production. The main aim of this study is to investigate these possibilities, to determine whether people speaking a non-native language pause with a foreign accent when reading aloud.

Previous studies on pausing behavior in non-native languages mostly concern fluency rather than foreign accent (e.g. [32,34–38]). Although a lack of fluency can contribute to recognizing speakers as non-native, fluency should be distinguished from accents, since proficient second language speakers might be highly fluent, but still possess a clear foreign accent [3,39–41]. Nonetheless, some tentative predictions can be drawn from fluency studies (see S3 Table): speakers tend to make more [32,42–47] and longer [32,46,47] (but also [32,42,43]) pauses when speaking their second language (L2) relative to their first language (L1). However, some reports find associations between speakers' L1 and L2 pausing behavior. When comparing L2 speakers to L1 speakers of the target language, it seems that highly proficient speakers adhere more closely to native-like pause patterns, whereas less proficient speakers make more and longer pauses than L1 speakers (L1 Korean, L2 English 8,L1 Russian, L2 English 34; see S4 Table). In addition, L1 and L2 pause frequency are correlated: while, overall, speakers make more pauses in their L2 than in their L1, speakers' L2 pausing behavior can be predicted if their L1 pausing behavior is known [47]. Finally, studies on the perception of fluency and accent can lead to conclusions about how pauses influence accents. For example, acoustic measures of fluency (including pause incidence and duration) have been shown to be predictors of accent ratings [3]. Still, perception of foreign accent was only weakly correlated with acoustic fluency measures, which is why, overall, accentedness and fluency can be regarded as two separate, partially independent concepts [3].

An important methodological issue implies that these previous studies cannot be taken as a clear evidence that pauses contribute to foreign accents in production: an intrinsic lack of stimulus control during free speech. Because pauses serve multiple functions, they can be influenced by many different variables, such as speech genre, cognitive load, or syntactic complexity of the utterances [34,48]. These diverse influencing factors make it a challenge to tease out the contributions of pauses to foreign accents in spontaneous speech (e.g., picture description tasks or spontaneous monologues). Differences in pause patterns may arise not due to foreign transference, but due to different sentence structures employed or speech styles adopted by the speakers, which cannot be controlled in spontaneous speech. Additional factors such as communicative intent, personal speaking style, or emotional involvement with the speech task [34,47–49] are equally difficult to control for. Most previous studies on second language pausing controlled for some of these potential factors but, due to their focus on second language fluency, did not control enough factors to reliably attribute foreign accents to pause differences.

Here we introduce an experimental procedure which integrates and modifies methods from previous studies, safeguarding against potential confounds, to focus on the specific contribution of pauses to foreign accents in speech production. We compared the pausing behavior of L2 speakers of English to the pausing behavior of L1 speakers of English, reading out the same scripted text at different speech rates. We measured multiple characteristics of pauses: pause-to-utterance ratio (total pause time in relation to total speaking time), the mean duration of pauses, the number of pauses that speakers made, and the positions in the written text at which they occurred.

By having all speakers read the same written text, we could exclude the influence of morpho-syntactic factors (e.g. word length, word structure, and syntactic structure) that might underlie previous findings of language differences in pausing. Also, we reasoned that the cognitive load involved in reading a text is reduced compared to free speech. Thus, by using written text as a prompt, we aimed to limit the occurrence of vocal hesitations resulting from a lack of L2 fluency to distinguish fluency from foreign accent per se.

To examine whether unusual reading conditions affect pausing behavior, we investigated speakers' pausing performance at three different reading speeds (casual, slow and fast speech rate), aiming to disentangle the role of cognitive load on the realization of pauses. We reasoned that the cognitive load should be lowest in casual reading speed because speakers encounter this speech tempo frequently in daily life and therefore have more possibilities to acquire native-like pause patterns [50]. Accents might preferentially surface in unnatural reading conditions, particularly in speeded reading where the cognitive load is higher and speakers might fall back on cognitive mechanisms developed for their native language. Also, testing the same individuals at different reading speeds offers a within-individual manipulation, helping to address individual differences in speaking and/or pausing styles.

We tested non-native speakers with two different L1 backgrounds, namely German or Serbo-Croatian L1. Typologically, both English and German belong to the Germanic language family, and are stress-based languages [51], whereas Serbo-Croatian is a Slavic language [52] and is not stress-based [25]. This selection of L1 speakers thus would allow comparison of pausing by English native speakers with those of L2 speakers with an L1 background more (German) or less (Serbo-Croatian) similar to English. However, this comparison was not included in the final analysis due to a low sample size of Serbo-Croatian speakers.

Summarizing, we used a standardized procedure to compare pause patterns (pause-to-utterance ratio, pause duration, pause number, pause positions) of native speakers of English and two non-native English speaker groups reading out the same text at three different speech rates. Previous work leads to several hypotheses and predictions.

According to the *No Contribution* hypothesis, pauses do not contribute to non-native speech production. This is predicted if the acquisition of pause patterns is simple during language acquisition, because pauses are perceptually highly salient (Matzinger, Ritt, Fitch, in prep) and pausing is articulatorily simple (compared to the articulation of other vocal elements like vowels, consonants, or intonation). Alternatively, L2-typical pause patterns might result from similar pause patterns in the speakers' L1 being transferred to their L2, or even reflect language-universal pausing behavior. By the *No Contribution* hypothesis, accents are caused by non-native realizations of linguistic features other than pauses. These non-native realizations might for example concern phonemes, word stress patterns or prosody [2,39]. Besides that, non-nativeness might be signaled by atypical gestures or turn-taking behavior. The *No Contribution* hypothesis predicts no differences in the pausing behavior of native and non-native speakers of English: speakers should pause at the same syntactic positions, equally often and for similar durations as native speakers. If pauses can be acquired easily or are language-independent, this hypothesis also predicts no interactions between nativeness and reading speed: L2 speakers should perform similarly to L1 speakers whether reading fast or slowly.

Alternatively, the *Pause Contribution* hypothesis postulates that pauses contribute to foreign accents. This might be due to a higher cognitive load (e.g. processing or memory constraints [18,53]) when speaking an L2 [8,54,55]. Alternatively, differences in the typical pausing behavior of speakers' L1 and L2 may make the typical pause characteristics of the L2 difficult to attain, because speakers' native pausing pattern overrides the learned non-native pattern. The *Pause Contribution* hypothesis predicts that pause patterns of L2 speakers will differ significantly from pause patterns of L1 speakers. If differences result from a higher cognitive load when speaking the L2, there should be more and longer pauses in L2 speakers than in L1 speakers. Also, this predicts an interaction between nativeness and reading speed. Differences between L1 and L2 pausing should be smaller in casual reading speed than in unnatural reading conditions (i.e. fast or slow speech), which pose a higher cognitive load because they are encountered and practiced less frequently. L2 speakers might therefore have had fewer chances to acquire them in a native-like manner and might fall back on non-native patterns instead. Differences are predicted to be higher in fast than in slow reading aloud, because reading rapidly should be more cognitively challenging than reading slowly.

Materials and methods

Target languages and participants

We obtained speech samples from 41 participants of three different first languages: English native speakers (13 participants; 7f; mean age: 35.2) and non-native English speakers with German (18 participants; 10f; mean age: 29.5) and Serbo-Croatian (10 participants; 6f; mean age: 25.9) as their first languages. Participants were university students or staff recruited individually at the University of Vienna. All non-native English participants were advanced learners of English, who did not have diagnosed reading or speaking difficulties, self-assessed themselves as being proficient in English (equivalent to CEFR level C1) and reported in post-experiment questionnaires (see S1 Appendix) to be concerned with English regularly both in the productive and receptive domain (e.g. in the university or work context, media exposure).

Nonetheless, and crucially, in a native language recognition test with our pool of speech samples, five English native speakers (m, mean age: 41.2) could still detect all non-native speakers due to their distinctive accents. This native language recognition test ensured that our non-native speakers qualified for the study: being identified as non-native in the accent recognition test suggests that certain features of native and non-native speech production differ. These might potentially include deviations in pause patterns.

The language recognition test was implemented using the software package PRAAT (Version 6.0.36, [56]). Raters listened to a speech sample of each participant reading out the target text (see below) in casual speech tempo. The task of the raters was to indicate if they believed the speakers to be native speakers of English, German or Serbo-Croatian. The raters controlled the timing, moving to the next speech sample as soon as they were sufficiently certain about their decision.

Although four German native speakers were misclassified as English native speakers by one or two of the raters each, the other raters correctly identified them as non-native. Also, one Serbo-Croatian native speaker was misclassified as an English native speaker by one of the raters, but correctly recognized as non-native by all other raters. The five raters correctly recognized all English native speakers, except that four English native speakers were classified as German L2 speakers by one rater each, and one of these English native speakers was additionally classified as a Serbo-Croatian L2 speaker by a second rater (see S5 Table). We concluded from these ratings that our speakers qualified for the analysis.

The study protocol was approved by the ethics board of the University of Vienna (reference number: #00333/00384). All subjects gave written informed consent in accordance with the Declaration of Helsinki.

Speech recordings

Speech samples were collected by recording participants reading out the English prose text *The boy who cried wolf* (see S2 Appendix), a fable frequently used for evaluating English pronunciation [57]. The recordings were made with a ZOOM Handy Recorder (H4n, ZOOM Corporation, Japan) either in a sound-proof room or in a quiet office environment. We recorded the participants reading out the text in three different speech tempi: fast, casually and slowly. Participants were instructed to read the text casually in the *casual* condition, to read as fast as they could in the *fast* condition and to read the text slowly (e.g., as if to a group of preschool children), in the *slow* condition. The order of the different tempo conditions was randomized for each participant. To elicit a natural reading style and minimize pauses resulting from hesitation due to unfamiliarity, participants were asked to read the text silently before recording in order to familiarize themselves with the text to avoid them being distracted by unfamiliar words or content during recording. In addition, before the actual recording, participants read the first sentence of the text aloud to make them comfortable reading aloud in the experimental setting, while the experimenter adjusted the signal recording levels.

Measurements and analyses

For determining pauses in the recordings of participants reading aloud, a pause was defined as a period of silence with a minimal duration of 0.1 seconds, most likely occurring for breathing, rhythmic or pragmatic reasons (the choice of this threshold is explained in more detail in Box 1).

Box 1. What is a pause?

Previous studies differ considerably in what they consider the lower durational threshold for classifying silent intervals in speech recordings as pauses (reviewed in [37,80]). These threshold values start as low as 5 ms (e.g. [25]) and range up to values as high as 400 ms [45,53,81]. 200 ms is a popular threshold for pauses in L2 speech [37,80]. The choice of a particular lower durational threshold is often determined by the type of pauses investigated in a particular study. For example, studies concerned with pauses resulting from a lack of fluency in an L2 tend to choose longer durational thresholds than studies investigating pauses in L1 everyday conversation. For our purpose, it was essential to sample all pauses, without excluding pauses below a lower durational limit. Still, we could not automatically classify all silent intervals in our recordings as pauses because silent intervals can be of multiple origins.

Silent intervals in speech recordings can occur because of pauses that fit the definition for our analysis, i.e. silent intervals resulting from breathing, rhythmic, structural or pragmatic reasons ("true pauses"), but also because of holds in stop consonants (for example /p/, /t/ or /k/) or very low amplitude schwas or fricatives (for example /f/, /s/ or /h/), i.e. silent intervals that are not considered as pauses ("phonetic silences"). Our automatic pause detection algorithm should only detect the former pauses, but not the latter ones. In order to determine the lower durational limit for the automatic detection mechanism, we determined the threshold below which no more true pauses occurred. For that, we analyzed the recordings of 6 speakers (pseudo-randomized, we ensured that there were 2 speakers of each language, one male and one female, 2 recordings for each condition, and 3 speakers in the sound-proof room and in the office). We automatically detected all silent intervals longer than 0.001 seconds (threshold -35 dB, minimum silent interval: 0.001 s) and then manually determined whether the detected silences were phonetic silences or true pauses. We then evaluated the distributional pattern of silences.

In these analyses we found that 90.39% (CIs: 80.89 and 99.9%) of the phonetic silences (n = 338) were shorter than 0.1 s, 9.29% (CIs: 0.44 and 18.14%) had durations between 0.1 and 0.2 s, and 0.32% (CIs: -0.5 and 1.13%) were longer than 0.2 s. In contrast, 93.14% (CIs: 86.09 and 100.1%) of the pauses (n = 132) were longer than 0.2 s, 6.86% (CIs: -0.18 and 13.91%) had durations between 0.1 and 0.2 s, and no true pauses were shorter than 0.1 s (Fig 1, Table 1). This led us to the conclusion to choose 0.1 s as a lower threshold for the automatic annotation of pauses to exclude most phonetic silences and include all true pauses. The remaining phonetic silences were deleted manually.

Pause measurement was performed using the software package PRAAT (Version 6.0.36, [56]). For that purpose, pauses were automatically annotated (Annotate → To TextGrid (silences); guidelines for settings: Silence threshold: -35.0 dB, Minimum silent interval duration: 0.1 s, Minimum sounding interval duration: 0.1 s). Additionally, all pauses were checked visually (in the oscillogram and spectrogram) and acoustically, and adjusted manually, in order to remove rarely occurring incorrectly identified pauses (e.g. holds in plosives, low amplitude fricatives; see Box 1) or to insert pauses that had not been automatically detected (for example, because of breathing noise). With a PRAAT script, the total duration of each reading, the number of pauses and the duration of individual pauses in each reading were extracted. We included all true silent pauses (see Box 1). Although we intended to include

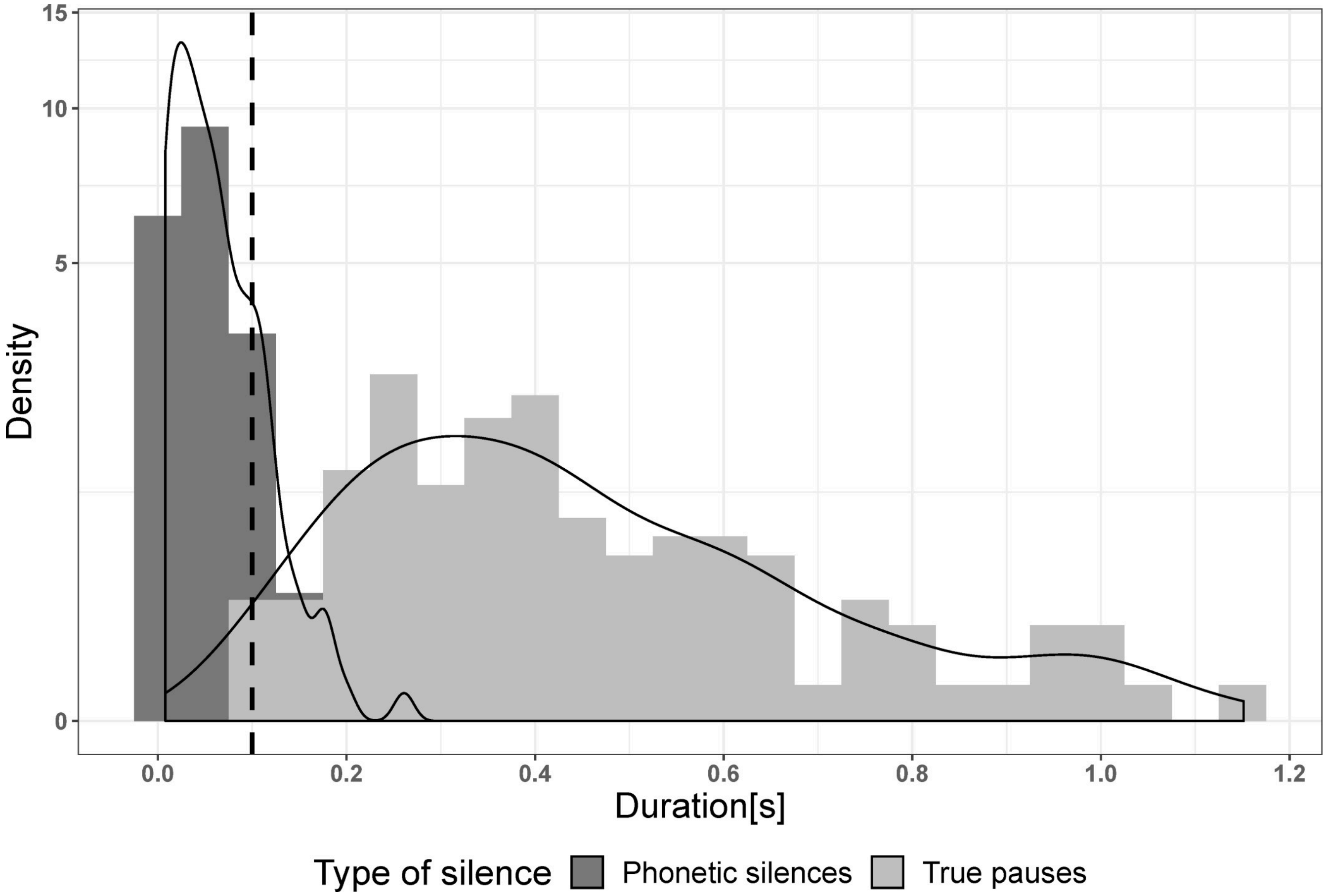


Fig 1. Distribution of phonetic silences and true pauses. Dashed line = threshold chosen for the subsequent automatic detection of pauses (0.1 s).

https://doi.org/10.1371/journal.pone.0230710.g001

Table 1. Mean pause duration and mean proportion of short (< 0.1 s), medium (0.1 < x < 0.2 s) and long (> 0.2 s) phonetic silences and true pauses with respective low and high confidence intervals (CIs).

	Duration [s]	Low CI	High CI	Short [%x]	Low CI	High CI	Medium [%]	Low CI	High CI	Long [%]	Low CI	High CI
Phonetic silences	0.047	0.033	0.061	90.39	80.89	99.9	9.29	0.44	18.14	0.32	-0.5	1.13
True pauses	0.404	0.301	0.506	0	0	0	6.86	-0.18	13.91	93.14	86.09	100.18

https://doi.org/10.1371/journal.pone.0230710.t001

filled pauses that contained a noise component resulting for example from breathing or from vocal hesitations such as "ehm" or "uh" [21,48,58,59], we did not find vocal hesitations in our data, most likely because participants read a scripted text and had familiarized themselves with the text before being recorded. Thus, the only filled pauses in our data are rarely occurring intervals containing obvious breathing noise.

We classified all pauses with regard to their position in the text. For each pause, we determined whether it occurred at a punctuation mark in the text (hereafter "marked pauses"; i.e. full stops, commas or quotation marks; no other punctuation marks occurred in the text), at an unmarked clause or phrase boundary (hereafter "unmarked pauses"; e.g. before a defining relative clause), or at any other position in the text. For the full text with the annotation of the pause categories see S2 Appendix.

We used linear and logistic mixed effects models to investigate the influence of reading tempo, native language and in-text position on the realization of pauses. Preliminary analyses did not reveal differences between native German and native Serbo-Croatian non-native speakers of English, so we lumped these two groups together for our analyses. Thus our analyses compared two groups, namely native and non-native speakers of English ("nativeness" factor). Still, in our plots, we present the data for all three native languages separately, in order to allow visual comparisons between them.

For our model predictors reading tempo and nativeness, we used deviation coding [60]. Reading tempo was coded as a continuous predictor (fast = -0.5, casual = 0, slow = +0.5), and nativeness was coded as a two-level factor (native = -0.5, non-native = +0.5).

To test whether total reading time, pause-to-utterance ratio and the duration of individual pauses were influenced by reading tempo and nativeness, we used linear mixed models [61] into which we entered these two predictors and an interaction term of the two as fixed effects. In order to reduce non-normality in the error structure of our models, the dependent variables total reading time and duration of individual pauses were log-transformed, because the optimal lambda for a Box-Cox transformation [62] was close to 0 in both cases, using the *boxcox* function of the *MASS* package [63]. The dependent variable pause-to-utterance ratio, which is proportional data bounded by 0 and 1, was logit-transformed, using the *logit* function of the *boot* package [64–67].

To investigate which factors influenced the probability of making a pause, we used a generalized linear mixed model [61] with binomial error structure and a logit link function. Each transition between two words represented a data point, and we determined for each of these transitions if a pause occurred there or not (similar to [38]). In total, this resulted in 26,456 data points. This number of data points can be explained as follows: 41 participants * 3 reading tempi * 215 word boundaries in the text = 26,445 data points. Additional 11 datapoints resulted from words that were not in the scripted text but that participants inserted spontaneously while reading. This yielded 26,456 data points in total, 2,750 of which were pauses. In this model, we included reading tempo, nativeness, in-text position and an interaction of reading tempo and nativeness as fixed effects.

All of our models included participant as a random intercept. For the three models testing the influence of reading tempo and nativeness on total reading time, pause-to-utterance ratio and the duration of individual pauses, our design, with one reading per tempo condition of each participant, did not allow us to accurately estimate random slopes. For the model testing the influence of reading tempo, nativeness and in-text position on the probability of making a pause, we ran an initial model that also included a random slope of position. However, this model did not converge. Thus, no random slopes are included in the models (but see [68–70]).

All models were fitted in R (version 3.5.1, [71]) and implemented in RStudio (version 1.1.456, [72]) using the *lmer* function of the *lme4* package [73].

For each linear mixed model, we visually inspected a qqplot and the residuals plotted against fitted values to check whether the assumptions of normally distributed and homogeneous residuals were fulfilled (using a function provided by Roger Mundry, Leipzig, Germany). These indicated no obvious deviations from normality or homoscedasticity.

Finally, we derived variance inflation factors (VIF, [74]) using the *vif* function of the R-packagae *car* applied to our models with the random effects excluded. They did not indicate collinearity to be an issue. We tested the significance of the respective full models as compared to the null models (comprising only the random intercept) by using a likelihood ratio test (R function *anova* with the argument test set to "Chisq", [75,76]). In all cases, parameters were estimated using maximum likelihood (rather than Restricted Maximum Likelihood, [77]) in order to allow for likelihood ratio tests. To obtain p-values for the individual effects, we conducted likelihood ratio tests comparing the full with respective reduced models ([69], R function *drop1*).

As indicators for the goodness-of-fit of our models, we follow [78] and report the marginal and conditional R^2 for each full model. The marginal R^2 (R^2_m) reveals the variance explained by the entirety of the fixed effects, and the conditional R^2 (R^2_c) reveals the variance explained by the entirety of the fixed and random effects. Thus, these measures can be taken as indicators for the effect size for the full models. We calculated R^2_m and R^2_c using the *r.squaredGLMM* function from the *MuMIn* package [79].

Results

Our instructions successfully elicited three desired reading aloud tempi: the full model for the total reading time was clearly significant compared to the null model (likelihood ratio test: $\chi^2 = 169.43$, df = 3, $p < 0.001$, effect size for the full model: $R^2_m = 0.72$, $R^2_c = 0.82$). Specifically, there was an effect of reading tempo on total reading time (likelihood ratio test: $\chi^2 = 163.67$, df = 1, $p < 0.001$), with reading time increasing from the fast to the slow condition. Furthermore, we found a significant main effect of nativeness on the total reading time (likelihood ratio test: $\chi^2 = 10.21$, df = 1, $p = 0.001$) with the total reading duration being higher in non-native speakers than in native speakers of English. There was no significant interaction effect of reading tempo and nativeness, i.e. non-native speakers did not change their reading durations differently in fast and slow tempo compared to native speakers (Table 2; Fig 2A; random effects: S8 Table).

We next explored how reading tempo and nativeness influenced the realization of pauses. In particular, we investigated the proportion of the total reading time that was devoted to pauses (pause-to-utterance ratio; Fig 2B), and how long the individual pauses were (individual pause duration; Fig 2C) at the different reading tempi in native and non-native speakers. We also explored how many pauses speakers made (number of pauses; Fig 2D) and how the probabilities that speakers made pauses at certain positions in the text were influenced by reading tempo and nativeness.

The effects of reading tempo and nativeness on the pause-to-utterance ratio

The full model for the pause-to-utterance ratio was significant compared to the null model (likelihood ratio test: $\chi^2 = 138.08$, df = 3, $p < 0.001$, effect size for the full model: $R^2_m = 0.58$,

Table 2. Results of the linear mixed model exploring the effects of reading tempo and nativeness on the total reading time (log-transformed). The table reports estimated model coefficients, standard errors (SE) and lower and upper confidence intervals (CI), χ^2 values of likelihood ratio tests and respective degrees of freedom (df) and p-values (P).

Term	Estimate	SE	lower CI	upper CI	χ^2	df	P
Intercept	4.07	0.03	4.02	4.13	[a]	[a]	[a]
Reading tempo[b]	0.51	0.04	0.43	0.60	163.67	1	< 0.001
NativenessNonNative[b]	0.11	0.03	0.05	0.18	10.21	1	0.001
Reading tempo:NativenessNonNative[(2)]	-0.02	0.05	-0.18	0.08	0.11	1	0.740

[a] not shown because of having a very limited interpretation

[b] Reading tempo and nativeness were deviation coded: reading tempo was coded as a continuous predictor (fast = -0.5, casual = 0, slow = +0.5), and nativeness was coded as a two-level factor (native = -0.5, non-native = +0.5); the indicated tests were obtained from likelihood ratio tests comparing the full with a reduced model lacking reading tempo, nativeness and the interaction, respectively.

https://doi.org/10.1371/journal.pone.0230710.t002

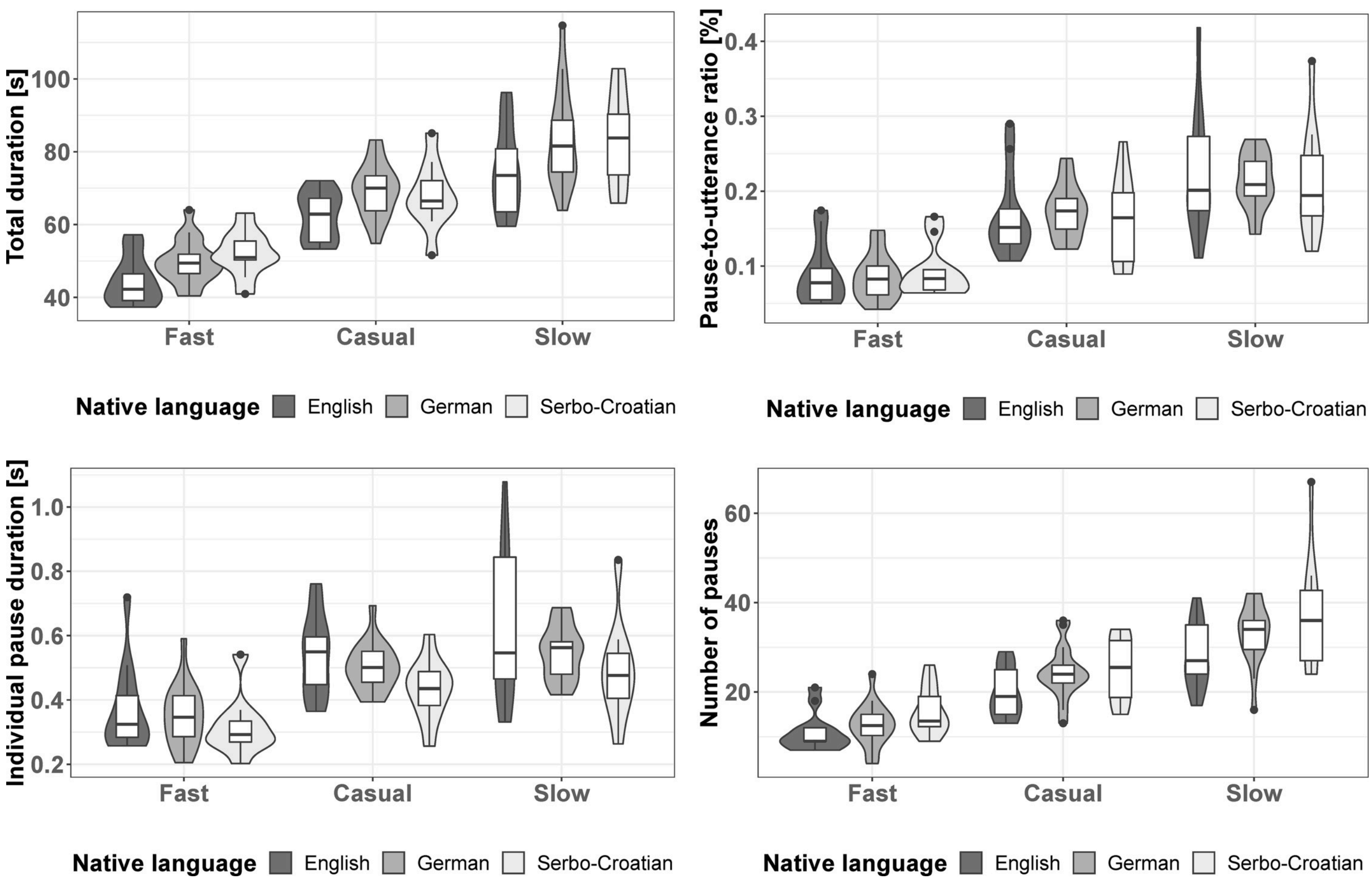


Fig 2. **a)** total duration of the readings in seconds, **b)** pause-to-utterance ratio in %, **c)** duration of individual pauses in seconds and **d)** number of pauses in each condition for each native language. The violin plots show median values (horizontal black lines) with first and third quartiles (lower and upper end of boxes), minimum and maximum values limited to values no more than 1.5 IQRs distant from the respective end of the box (lower and upper end of vertical black lines) and outliers (black dots). The area around each box indicates the distribution of the data.

https://doi.org/10.1371/journal.pone.0230710.g002

R^2_c = 0.80). More specifically, reading tempo had a significant effect on the pause-to-utterance ratio (likelihood ratio test: χ^2 = 137.94, df = 1, p < 0.001). In contrast, we did not find a significant effect of nativeness on the pause-to-utterance ratio (likelihood ratio test: χ^2 = 0.002, df = 1, p = 0.965). Also, the effect of the interaction of reading tempo and nativeness was non-significant (likelihood ratio test: χ^2 = 0.14, df = 1, p = 0.709). Non-native speakers thus spent a similar amount of time on pauses as native speakers.

Averaged over native and non-native speakers, the mean pause-to-utterance ratio was 22.62% for slow, 14.35% for casual, and 8.76% for fast reading aloud (Table 3; Fig 2B; random effects: S9 Table). We used the following formulae for the back-transformation from the logit-transformed model estimates (Table 3): odds = exp(– 1.79 + 1.11 * reading tempo + 0.01 * nativeness– 0.05 * reading tempo * nativeness, and y = odds/(1+odds). Reading tempo and nativeness were deviation coded: reading tempo was coded as a continuous predictor (fast = -0.5, casual = 0, slow = +0.5), and nativeness was coded as a two-level factor (native = -0.5, non-native = +0.5). The respective values for fast, casual and slow reading speeds were inserted into the formulae to calculate the mean pause-to-utterance ratios. To get values averaged for native and non-native speakers, we inserted 0 for nativeness.

Table 3. Results of the linear mixed model exploring the effects of reading tempo and native language on pause-to-utterance ratio (logit-transformed). The table reports estimated model coefficients, standard errors (SE) and lower and upper confidence intervals (CI), χ^2 values of likelihood ratio tests and respective degrees of freedom (df) and p-values (P).

Term	Estimate	SE	lower CI	upper CI	χ^2	df	P
Intercept	-1.79	0.09	-1.96	-1.61	[a]	[a]	[a]
Reading tempo[b]	1.11	0.10	0.91	1.32	137.94	1	< 0.001
NativenessNonNative[b]	0.01	0.10	-0.20	0.21	0.002	1	0.965
Reading tempo: NativenessNonNative(2)	-0.05	0.12	-0.29	0.20	0.14	1	0.709

[a] not shown because of having a very limited interpretation

[b] Reading tempo and nativeness were deviation coded: reading tempo was coded as a continuous predictor (fast = -0.5, casual = 0, slow = +0.5), and nativeness was coded as a two-level factor (native = -0.5, non-native = +0.5); the indicated tests were obtained from likelihood ratio tests comparing the full with a reduced model lacking reading tempo, nativeness and the interaction, respectively.

https://doi.org/10.1371/journal.pone.0230710.t003

The effects of reading tempo and nativeness on the duration of individual pauses

The full model for the individual pause durations was significant compared to the null model (likelihood ratio test: $\chi^2 = 89.46$, df = 3, $p < 0.001$, effect size for the full model: $R^2_m = 0.37$, $R^2_c = 0.73$). We found that there was a significant effect of reading tempo on the duration of individual pauses (likelihood ratio test: $\chi^2 = 87.53$, df = 1, $p < 0.001$). Contrastingly, we did not find a significant effect of nativeness on pause duration (likelihood ratio test: $\chi^2 = 1.73$, df = 1, p = 0.188). Likewise, the effect of the interaction of reading tempo and nativeness was non-significant (likelihood ratio test: $\chi^2 = 0.21$, df = 1, p = 0.651), which indicates that native and non-native speakers of English altered the duration of their pauses similarly in different reading tempi.

The mean duration of individual pauses, averaged over native and non-native speakers, was 0.60 s in slow, 0.47 s in casual, and 0.37 s in fast reading aloud (Table 4; Fig 2C; random effects: S8 Table). We used the following formula for the back-transformation from the log-transformed model estimates (Table 4): $y = \exp(-0.75 + 0.49 * \text{reading tempo} - 0.10 * \text{nativeness} - 0.04 * \text{reading tempo} * \text{nativeness})$. Reading tempo and nativeness were deviation coded: reading tempo was coded as a continuous predictor (fast = -0.5, casual = 0, slow = +0.5), and nativeness was coded as a two-level factor (native = -0.5, non-native = +0.5). The respective values for fast, casual and slow speech were inserted into the formula to calculate the mean durations. To get values averaged for native and non-native speakers, we inserted 0 for nativeness.

Table 4. Results of the linear mixed model exploring the effects of reading tempo and native language on the duration of individual pauses (log-transformed). The table reports estimated model coefficients, standard errors (SE) and lower and upper confidence intervals (CI), χ^2 values of likelihood ratio tests and respective degrees of freedom (df) and p-values (P).

Term	Estimate	SE	lower CI	upper CI	χ^2	df	P
Intercept	-0.75	0.06	-0.87	-0.63	[a]	[a]	[a]
Reading tempo[b]	0.49	0.07	0.36	0.62	87.53	1	< 0.001
NativenessNonNative[b]	-0.10	0.07	-0.24	0.05	1.73	1	0.188
Reading tempo: NativenessNonNative(2)	-0.04	0.08	-0.19	0.12	0.25	1	0.651

[a] not shown because of having a very limited interpretation

[b] Reading tempo and nativeness were deviation coded: reading tempo was coded as a continuous predictor (fast = -0.5, casual = 0, slow = +0.5), and nativeness was coded as a two-level factor (native = -0.5, non-native = +0.5); the indicated tests were obtained from likelihood ratio tests comparing the full with a reduced model lacking reading tempo, nativeness and the interaction, respectively.

https://doi.org/10.1371/journal.pone.0230710.t004

Table 5. Results of the logistic regression model exploring the effects of reading tempo, native language and in-text position on the occurrence frequency of pauses. The table reports estimated model coefficients, standard errors (SE) and lower and upper confidence intervals (CI), χ^2 values of likelihood ratio tests and respective degrees of freedom (df) and p-values (P).

Term	Estimate	SE	lower CI	upper CI	χ^2	df	P
Intercept	0.60	0.17	0.27	0.94	[a]	[a]	[a]
PositionUnmarkedPhraseBoundary[b]	-2.73	0.08	-2.89	-2.57	9759.82	1	< 0.001[c]
PositionOtherWordBoundary[b]	-6.06	0.09	-6.25	-5.88			
Reading tempo[b]	2.20	0.15	1.91	2.48	906.75	1	< 0.001
NativenessNonNative[b]	0.54	0.20	0.14	0.94	7.05	1	0.008
Reading tempo: NativenessNonNative[b]	0.27	0.17	-0.06	0.60	2.59	1	0.11

[a] not shown because of having a very limited interpretation

[b] In-text position was dummy coded with the marked position being the respective reference category. Reading tempo and nativeness were deviation coded: reading tempo was coded as a continuous predictor (fast = -0.5, casual = 0, slow = +0.5), and nativeness was coded as a two-level factor (native = -0.5, non-native = +0.5); the indicated tests were obtained from likelihood ratio tests comparing the full with a reduced model lacking in-text position, reading tempo, nativeness and the interaction between reading tempo and nativeness, respectively.

[c] This is the overall effect of in-text position on the occurrence frequency of pauses. Considering the differences between the individual levels, the model revealed that at unmarked phrase boundaries (z = -33.49, p < 0.001) and at other word boundaries (z = -65.17, p < 0.001), participants paused significantly less than at punctuation marks.

https://doi.org/10.1371/journal.pone.0230710.t005

The effect of reading tempo, nativeness and position in the text on the occurrence frequency of pauses

The full model for the occurrence frequency of pauses was significant compared to the null model (likelihood ratio test: $\chi^2 = 10186$, df = 5, $p < 0.001$, effect size for the full model: $R^2_m = 0.32$, $R^2_c = 0.34$). We found significant effects of reading tempo (likelihood ratio test: $\chi^2 = 906.75$, df = 1, $p < 0.001$), of nativeness (likelihood ratio test: $\chi^2 = 7.05$, df = 1, p = 0.008) and of in-text position (likelihood ratio test: $\chi^2 = 9759.82$, df = 1, $p < 0.001$) on the occurrence frequency of pauses. Non-native speakers made more pauses than native speakers, and people made more pauses the more slowly they read aloud. Also, people made more pauses at punctuation marks than at unmarked phrase boundaries and at other positions in the text. However, the effect of the interaction of reading tempo and nativeness was non-significant (likelihood ratio test: $\chi^2 = 2.59$, df = 1, p = 0.11), which indicates that native and non-native speakers of English altered the occurrence frequency of their pauses similarly in different reading aloud tempi (Table 5). Regarding our random intercept of participant, the estimated standard deviation among participants was 0.56 (S9 Table). This is smaller than the magnitude of effects of in-text position and reading tempo, but similar to the magnitude of the effect of nativeness (cf. Table 5). This indicates that the influence of participant variability and nativeness are comparable.

Native and non-native speakers' predicted probabilities of making a pause in fast, casual and slow reading at punctuation marks, unmarked phrase boundaries and other word boundaries are given in Table 6. The corresponding partial effects of our model are shown in Fig 3. The overall numbers of pauses in each reading tempo and each native language are displayed in Fig 2D.

Discussion

Our study evaluated whether pauses contributed to the speakers' non-native L2 production by examining whether native and non-native speakers of English differed in their pausing behavior when reading aloud at different speech rates. Although there were some rare misclassifications in our native language recognition test, overall, all non-native speakers were identified as

Table 6. Predicted probabilities (in %) of making a pause in fast, casual or slow reading at punctuation marks, unmarked phrase boundaries or other positions in the text for native and non-native speakers of English.

	Native[a]			Non-Native[a]		
	Fast	**Casual**	**Slow**	**Fast**	**Casual**	**Slow**
Punctuation mark	33.20	58.19	79.58	42.69	70.50	88.46
Phrase boundary	3.15	8.34	20.31	4.65	13.51	33.39
Other word boundary	0.12	0.32	0.90	0.17	0.55	1.76

[a] Values are derived from the estimates of our logistic regression model (see Table 5). We used the following formulae for the back-transformation from the logit-transformed model estimates: odds = exp(0.60–2.73 * unmarked phrase boundary– 6.06 * other word boundary + 2.20 * reading tempo + 0.54 * nativeness + 0.27 * reading tempo * nativeness, and y = odds/(1+odds). In-text position was dummy coded, with marked boundaries serving as the reference category. Reading tempo and nativeness were deviation coded: reading tempo was coded as a continuous predictor (fast = -0.5, casual = 0, slow = +0.5), and nativeness was coded as a two-level factor (native = -0.5, non-native = +0.5). The respective values were inserted into the formulae to calculate the predicted probabilities of pause occurrence.

https://doi.org/10.1371/journal.pone.0230710.t006

non-native by native English raters, and thus had clearly recognizable non-native accents. However, our findings suggest that non-native pause patterns contributed little to the production of these non-native accents, at least for our relatively proficient speakers (see below). This supports the *No Contribution* hypothesis.

First, native and non-native speakers had similar pause-to-utterance ratios. Second, the durations of their pauses were similar (in line with 25,33,but in contrast to 46). Third, for native and non-native speakers, reading tempo had a similar significant influence on all of the pause characteristics measured, and there were no interactions between nativeness and

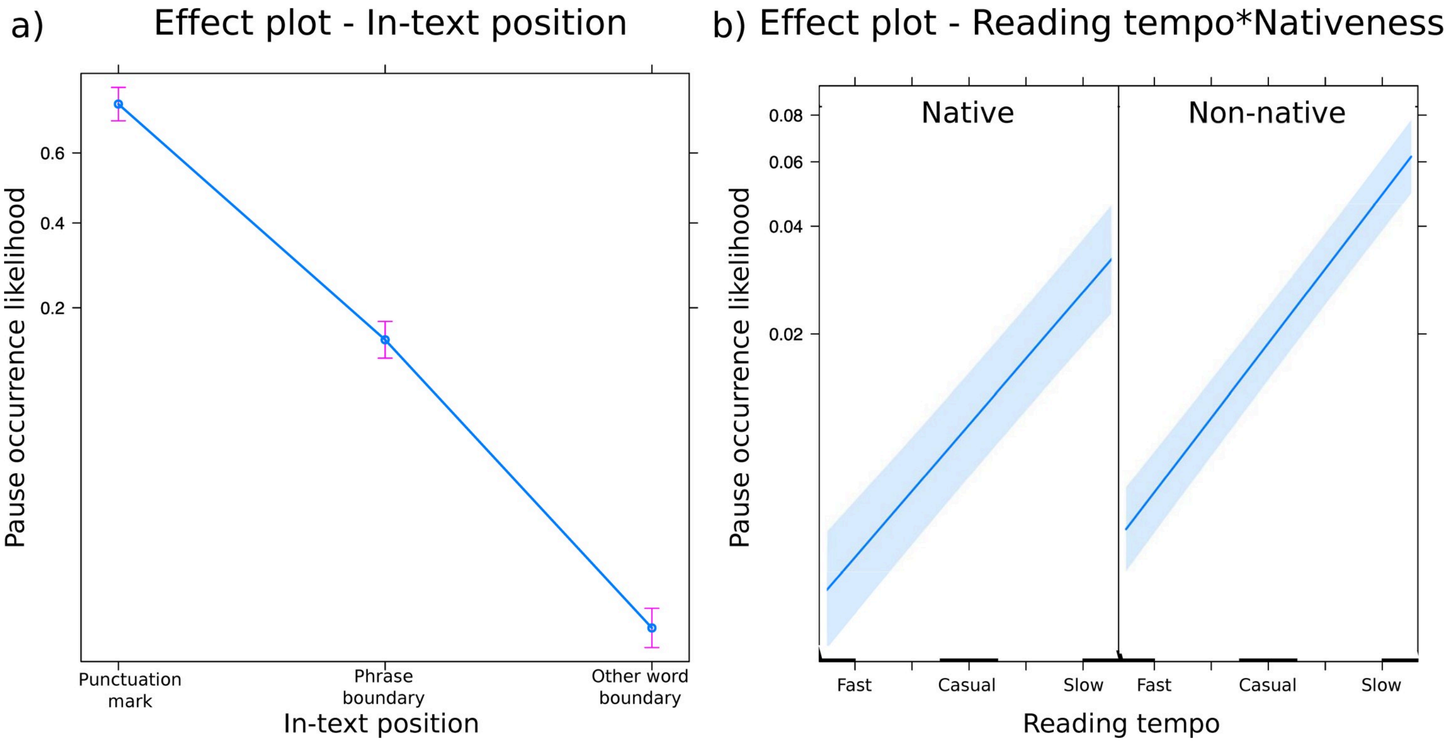


Fig 3. a) effect display of the significant main effect of in-text position, **b)** effect display of the significant main effects of reading tempo and nativeness. The effect of the interaction of reading tempo and nativeness was non-significant. The y-axis displays the probability of occurrence of a word after a pause. Error bars and shaded areas around the estimated effects represent 95% confidence intervals.

https://doi.org/10.1371/journal.pone.0230710.g003

reading rate. With an increasing reading tempo, native and non-native speakers made fewer and shorter pauses, and with a decreasing reading tempo, they made more and longer pauses. Thus, with changing reading tempo, native and non-native speakers altered both the duration and the occurrence likelihood of pauses in a highly similar way. During fast reading, native and non-native speakers' probabilities of making a pause were below 5% at unmarked phrase boundaries and below 0.2% at other unmarked positions in the text. Thus, almost only pauses at punctuation marks remained, suggesting that the visually salient punctuation marks help readers to structure their vocal output in similar ways.

Pause-to-utterance ratio changed with changing reading tempo (cf. similar findings in speakers of Dutch, French and Italian; [26]), indicating that in fast reading, pauses were reduced more compared to casual reading and in slow reading, pauses increased relative to casual reading. A potential explanation is that when reducing pauses, there is less information loss compared to reducing articulated sounds. If articulated sounds were deleted or shortened too much, words would be distorted to intelligibility and semantic content would be lost. Similarly, if segments were lengthened to an unnatural extent they would be difficult to produce and perceive. These factors apply much less to shortened or lengthened pauses, and reducing or increasing pause length is articulatorily simpler than varying the duration of vocal segments. This may explain the similar results regarding reading tempo for native and non-native speakers.

The only significant difference in pauses that may contribute to sounding foreign is the higher likelihood of having a pause in L2 speech: non-native speakers made more pauses than native speakers (in contrast to [25]). Although our non-native participants were highly proficient in English, they might still need more time for cognitive processing when speaking their L2 [46,82]. This might be reflected in their higher likelihood of occurrence of pauses (but interestingly not in longer durations of pauses). However, the reason speakers need more processing time for the foreign language might be rooted not in difficulties in adhering to target language pause patterns, but to other aspects of the L2, such as difficulties pronouncing L2-typical sounds. Further evidence that speakers need more time for cognitive processing in their L2 than in their L1 is that non-native speakers had longer reading durations than native speakers (cf. [83]). To rule out the possibility that non-native speakers have a slower reading speed in general, the reading durations in their respective first languages would need to be addressed in a follow-up study. Also, this would shed light on the influence of individual participants' reading proficiency on our results (see below).

There are several possible explanations why, overall, our non-native speakers did not appear to produce pauses 'with a foreign accent'. Pauses may be easier to acquire than other aspects of language because they are perceptually salient (Matzinger, Ritt, Fitch, in prep) and pausing is articulatorily easy, relative to phonemes or intonation patterns. Alternatively, pauses might not contribute to non-native speech production because the pause patterns in the speakers' L1 and L2 might be similar and speakers simply transfer their L1 pause patterns to their L2. Such a transfer might imply that pauses result from very general cognitive mechanisms [53] and thus have a more universal character than other aspects of language. However, to substantiate this hypothesis, non-native English speakers of a typologically diverse set of languages, including many other native languages than German and Serbo-Croatian, would need to be tested in a larger-scale study. If such L2 speakers still pause with a native-like accent when speaking English, this would be evidence for a language universal character of pause patterns.

If pauses result from basic cognitive mechanisms, L2 speakers should also pause with a non-native accent in other second languages than English.

Although our study controls for many aspects necessary to investigate foreign speech production, there are some aspects that it does not address, but that could be addressed in potential follow-up experiments. Our study tested the role of pauses of highly proficient L2 speakers. Certainly, L2 proficiency might have an influence on the realization of pauses [8,34]. L2-typical pause patterns might be acquired rapidly (and should thus not contribute to non-native accents in any proficiency level), slowly but still more quickly than e.g. phonemes (and should thus only contribute to non-native accents in beginners), or not at all (and should thus contribute to non-native accents in any proficiency level). Therefore, testing L2 speakers of different proficiency levels in a follow-up experiment might reveal more about the cognitive constraints associated with speaking an L2, which might contribute to non-native speech production. Our present finding that pauses contribute little to the acoustic peculiarities of L2 production would be further corroborated if a comparable study with less proficient non-native speakers yielded similar results.

Furthermore, we caution that our relatively small sample size of 41 participants, although clearly adequate to reveal multiple statistically significant effects, might potentially be inadequate to reveal more subtle differences of smaller effect size. Thus, like any null result, our "no difference" findings should be viewed with some caution. However, the considerable time and effort required to derive, annotate and manually check pause data (more than 26,000 data points of which more than 2,700 were pauses comprise our current dataset) would remain a challenge for gathering much larger samples of participants.

We only included academically educated participants, from which a similarly high level of reading proficiency could be assumed. Testing people with a high reading proficiency might have contributed to the fact that, overall, we found similar pause patterns in native and non-native speakers. Highly proficient readers might be able to override challenges in foreign speech production when reading, but not when speaking freely (see below). Less proficient readers might not be able to mask difficulties in non-native speech production when reading, which might result in different pause patterns between native and non-native speakers. A potential follow-up study could explicitly test participants' reading proficiency and include it as a predictor for pause patterns.

The text that the participants read out in our study contained punctuation marks. Punctuation marks are salient visual cues that might prime participants' pause patterns [84]. We used a text containing punctuation marks to make the procedure as close to real-life reading situations as possible. How much punctuation contributes to adopting the pause patterns of an L2 in read speech could easily be tested in a follow-up study with texts without punctuation marks. Similar results using texts without punctuation marks could evaluate the possibility that the realization of pauses is determined by basic cognitive processes, because in such cases participants would not be primed by visual cues.

Results of our study on read texts might not be entirely transferrable to spontaneous speech, because pauses in reading aloud might also reflect reading difficulties or spelling-to-sound difficulties. In contrast, pauses in spontaneous speaking might reflect difficulties in conceptualizing the message or in linguistic formulation. Still, we argue that testing speakers during reading aloud has ecological relevance because it occurs in several real-life contexts, such as when reading aloud to children or in the (language) classroom. Especially during second language learning, reading aloud is a commonly practiced exercise [85]. Furthermore, reading is ideal for our purposes since it captures difficulties in articulation which are highly relevant for non-native speech production. Further, reading a complete story also captures pauses in longer stretches of speech, as opposed to reading or repeating isolated sentences (e.g. [8]). Thus, we feel that our paradigm adds a rigorous and useful new method to the existing literature on spontaneous speech.

One crucial point that our procedure cannot address is whether it is a transfer from L1 pause patterns that either hinders (because L1 and L2 pause patterns are different) or facilitates (because L1 and L2 pause patterns are similar) the acquisition of L2 typical pause patterns. Testing similarities or differences of different native languages' pause patterns is almost impossible: even if speakers of different L1s are tested on similar tasks, such as reading out texts matched in syntax and content, translations can never be fully identical in syntactic structure or small nuances of content. Such minute differences might already shape pause patterns in a way that makes it difficult to determine these potential differences' contribution to L2 accents. Nonetheless, testing the same speakers in their L1 and L2 would be useful to address the effect of individual speaking style on pause patterns [47].

Conclusions and future work

We asked native and non-native speakers to read the same English text, thus excluding potential L1-specific morpho-syntactic factors from influencing the vocal output. We found that speakers inserted pauses into the reading stream in similar ways and at similar locations, and changed this pattern in similar ways at different reading tempi, regardless of their native language. The only difference between pause patterns in native and non-native speakers was that the non-native speakers made more pauses than the native speakers. This might reflect cognitive processing constraints in an L2 that result from other aspects of the L2 than pausing behavior per se. Overall, we conclude that in reading aloud, the influence of nativeness on the realization of pauses is marginal, suggesting that pauses play little role in the production of foreign accents in this context

Supporting information

S1 Table. Results of cross-linguistic studies suggesting that the numbers and durations of pauses in different languages are similar.
(DOCX)

S2 Table. Results of cross-linguistic studies suggesting that the numbers and durations of pauses in different languages are different.
(DOCX)

S3 Table. Results of studies on the numbers and durations of pauses during L2 speech. Results in the table concern comparisons between pauses in speakers' L2s and these speakers' L1s (as opposed to L1 speakers of the target L2).
(DOCX)

S4 Table. Results of studies on the numbers and durations of pauses during L2 speech. Results in the table concern comparisons between pauses in speakers' L2s and L1 speakers of the target L2.
(DOCX)

S5 Table. Native language recognition ratings.
(DOCX)

S6 Table. Estimated variance components and standard deviations for the random intercept of participant of the full model exploring the effects of reading tempo, and nativeness on the total reading time.
(DOCX)

S7 Table. Estimated variance components and standard deviations for the random intercept of participant of the full model exploring the effects of reading tempo and nativeness on pause-to-utterance ratio.
(DOCX)

S8 Table. Estimated variance components and standard deviations for the random intercept of participant of the full model exploring the effects of reading tempo, and nativeness on the duration of individual pauses.
(DOCX)

S9 Table. Estimated variance components and standard deviations for the random intercept of participant of the full model exploring the effects of reading tempo, nativeness and in-text position on the occurrence frequency of pauses.
(DOCX)

S1 Appendix. Post-experiment questionnaire.
(DOCX)

S2 Appendix. Annotated text *The boy who cried wolf*.
(DOCX)

Acknowledgments

We thank Kamil Kaźmierski, Andreas Baumann, and two anonymous reviewers for helpful comments on previous versions, and Klaus Hofmann and Magdalena Schwarz for helpful discussions.

Author Contributions

Conceptualization: Theresa Matzinger, W. Tecumseh Fitch.

Investigation: Theresa Matzinger.

Methodology: Theresa Matzinger, W. Tecumseh Fitch.

Supervision: Nikolaus Ritt, W. Tecumseh Fitch.

Writing – original draft: Theresa Matzinger.

Writing – review & editing: Nikolaus Ritt, W. Tecumseh Fitch.

References

1. Anderson-Hsieh J, Johnson R, Koehler K. The Relationship Between Native Speaker Judgments of Nonnative Pronunciation and Deviance in Segmentais, Prosody, and Syllable Structure. Lang Learn. 1992; 42(4):529–55.
2. Idemaru K, Wei P, Gubbins L. Acoustic Sources of Accent in Second Language Japanese Speech. Lang Speech. 2019; 62(2):333–57. https://doi.org/10.1177/0023830918773118 PMID: 29764295
3. Pinget AF, Bosker HR, Quené H, de Jong NH. Native speakers' perceptions of fluency and accent in L2 speech. Lang Test. 2014; 31(3):349–65.
4. Bransford JD, Brown AL, Cocking RR. How people learn—brain, mind, experience, and school. Washington D.C.: National Academy Press; 1999.
5. Major RC. Foreign Accent—The Ontogeny and Phylogeny of Second Langugae Phonology. Mahwah: Lawrence Erlbaum Associates; 2001.
6. Flege JE. Second Language Speech Learning: Theory, Findings, and Problems. In: Strange W, editor. Speech Perception and Linguistic Experience: Issues in Cross-Language Reserach. Timonium, MD: York Press; 1995. p. 233–77.

7. Chun DM. Discourse intonation in L2: From theory and research to practice. Amsterdam: John Benjamins Publishing; 2002.

8. Trofimovich P, Baker W. Learning Second Language Suprasegmentals: Effect of L2 Experience on Prosody and Fluency Characteristics of L2 Speech. Stud Second Lang Acquis. 2006; 28(1):1–30.

9. Iverson P, Kuhl PK, Akahane-Yamada R, Diesch E, Tohkura Y, Kettermann A, et al. A perceptual interference account of acquisition difficulties for non-native phonemes. Cognition. 2003; 87:B47–57. https://doi.org/10.1016/s0010-0277(02)00198-1 PMID: 12499111

10. König E, Gast V. Understanding English-German contrasts. 3rd ed. Berlin: Erich Schmidt Verlag; 2012.

11. Gibbon D. Intonation in German. In: Hirst D, Di Cristo A, editors. Intonation systems: a survey of twenty languages. Cambridge: Cambridge University Press; 1999. p. 78–95.

12. Mennen I, Schaeffler F, Dickie C. Second Language Acquisition of Pitch Range in German Learners of English. Stud Second Lang Acquis. 2014; 36(2):303–29.

13. Mennen I, Schaeffler F, Docherty G. Cross-language differences in fundamental frequency range: A comparison of English and German. J Acoust Soc Am. 2012; 131(3):2249–60. https://doi.org/10.1121/1.3681950 PMID: 22423720

14. Hockett CF. The Origin of Speech. Sci Am. 1960; 203:88–111. https://doi.org/10.1038/scientificamerican1260-88

15. Levinson SC. Pragmatics, Universals in. In: Hogan PC, editor. The Cambridge encyclopedia of the language sciences. New York: Cambridge University Press; 2011. p. 654–7.

16. Stivers T, Enfield NJ, Brown P, Englert C, Hayashi M, Heinemann T, et al. Universals and cultural variation in turn-taking in conversation. Proc Natl Acad Sci. 2009; 106(26):10587–92. https://doi.org/10.1073/pnas.0903616106 PMID: 19553212

17. Evans N, Levinson SC. The myth of language universals: Language diversity and its importance for cognitive science. Behav Brain Sci. 2009;(32):429–92.

18. Seifart F, Strunk J, Danielsen S, Hartmann I, Pakendorf B, Wichmann S, et al. Nouns slow down speech across structurally and culturally diverse languages. Proc Natl Acad Sci [Internet]. 2018; 115(22):5720–5. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1800708115PMID: 29760059

19. Goldman-Eisler F. Psycholinguistics: experiments in spontaneous speech. London: Academic Press; 1968.

20. Nooteboom S. The Prosody of Speech: Melody and Rhythm. In: Hardcastle WJ, Laver J, Gibbon FE, editors. The handbook of phonetic sciences. Oxford: Blackwell; 1997. p. 640–73.

21. Oliveira M. The role of pause occurrence and pause duration in the signaling of narrative structure. Adv Nat Lang Process [Internet]. 2002;43–51. Available from: http://link.springer.com/chapter/10.1007/3-540-45433-0_7

22. Swerts M. Prosodic features at discourse boundaries of different strength. J Acoust Soc Am. 1997; 101(1):514–21. https://doi.org/10.1121/1.418114 PMID: 9000742

23. Yang X, Shen X, Li W, Yang Y. How listeners weight acoustic cues to intonational phrase boundaries. PLoS One. 2014; 9(7):1–9.

24. Huensch A, Tracy-Ventura N. Understanding second language fluency behavior: the effects of individual differences in first language fluency, cross-linguistic differences, and proficiency over time. Appl Psycholinguist. 2016;1–31.

25. Smiljanic R, Bradlow AR. Production and perception of clear speech in Croatian and English. J Acoust Soc Am. 2005; 118:1677–88. https://doi.org/10.1121/1.2000788 PMID: 16240826

26. Demol M, Verhelst W, Verhoeve P. The duration of speech pauses in a multilingual environment. Proc Annu Conf Int Speech Commun Assoc INTERSPEECH. 2007; 1(1):117–20.

27. Yang L. Duration and pauses as boundary-markers in speech: a cross-linguistic study. In: Proceedings of Interspeech 2007 [Internet]. 2007. p. 458–61. Available from: http://www.speech.kth.se/prod/publications/files/100444.pdf

28. Grosjean FE, Deschamps A. Analyse contrastive des variables temporelles de l'anglais et du francais: vitesse de parole et variables composantes, phénomènes d'hésitation. Phonetica. 1975; 31:144–84.

29. Holmes VM. A crosslinguistic comparison of the production of utterances in discourse. Cognition. 1995; 54:169–207. https://doi.org/10.1016/0010-0277(94)00635-x PMID: 7874876

30. De Jong NH, Steinel MP, Florijn A, Schoonen R, Hulstijn JH. Linguistic skills and speaking fluency in a second language. Appl Psycholinguist. 2013; 34(5):893–916.

31. Johnson TH, O'Connell DC, Sabin EJ. Temporal analysis of English and Spanish narratives. Bull Psychon Soc. 1979; 13(6):347–50.

32. Trouvain J, Fauth C, Möbius B. Breath and Non-breath Pauses in Fluent and Disfluent Phases of German and French L1 and L2 Read Speech. In: Speech Prosody (SP8) [Internet]. 2016. p. 31–5. Available from: http://www.ifcasl.org/docs/Trouvain_Fauth_Moebius_2016.pdf

33. Campione E, Véronis J. A large-scale multilingual study of pause duration. Speech Prosody 2002 Proc 1st Int Conf Speech Prosody [Internet]. 2002;199–202. Available from: http://www.isca-speech.org/archive/sp2002/sp02_199.html

34. Riazantseva A. Second Language Proficience and Pausing: A Study of Russian Speakers of English. Stud Second Lang Acquis [Internet]. 2001; 23(4):497–526. Available from: http://www.journals.cambridge.org/abstract_S027226310100403X

35. Derwing TM, Munro MJ, Thomson RI, Rossiter MJ. The relationship between L1 fluency and L2 fluency development. Stud Second Lang Acquis. 2009; 31(4):533–57.

36. Rose R. Temporal variables in first and second language speech and perception of fluency. ICPhS 2015 Proc 18th Int Congr Phonetic Sci [Internet]. 2015;1–5. Available from: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0405.pdf

37. De Jong NH, Bosker HR. Choosing a threshold for silent pauses to measure second language fluency. DiSS 2013 Proc 6th Work Disfluency Spontaneous Speech [Internet]. 2013;17–20. Available from: http://www.isca-speech.org/archive/diss_2013/dis6_017.html

38. De Jong NH. Predicting pauses in L1 and L2 speech: the effects of utterance boundaries and word frequency. Int Rev Appl Linguist Lang Teach. 2016; 54(2):113–32.

39. Derwing T, Munro M. Accent, intelligibility, and comprehensibility: Evidence from Four L1s. Stud Second Lang Acquis. 1997; 19(1):1–16.

40. Derwing TM, Munro MJ, Thomson RI. A longitudinal study of ESL learners' fluency and comprehensibility development. Appl Linguist. 2008; 29(3):359–80.

41. Thomson RI. Fluency. In: Reed M, Levis JM, editors. The Handbook of English Pronunciation. Chichester: John Wiley & Sons; 2018. p. 209–26.

42. Raupach M. Temporal variables in first and second language speech and perception of fluency. In: Dechert HW, Raupach M, editors. Temporal variables in speech: Studies in honour of Frieda Goldman-Eisler [Internet]. The Hague: Mouton; 1980. p. 263–70. Available from: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0405.pdf

43. Deschamps A. The syntactical distribution of pauses in English spoken as a second language by French students. In: Dechert HW, Raupach M, editors. Temporal Variables in Speech: Studies in honour of Frieda Goldman-Eisler. The Hague: Mouton; 1980. p. 255–62.

44. Isarankura S. Variability of Pause Patterns in English Read Speech of Thai EFL Learners. J Educ Soc Res [Internet]. 2013; 3(7):346–54. Available from: http://www.mcser.org/journal/index.php/jesr/article/view/971

45. Tavakoli P. Pausing patterns: Differences between L2 learners and native speakers. ELT J. 2011; 65 (1):71–9.

46. Kolly M-J, Leemann A, Boula P, Mareüil D, Dellwo V. Speaker-idiosyncrasy in pausing behavior: evidence from a cross-linguistic study. Proc Int Congr Phonetic Sci 2015 Glas. 2015;1–5.

47. De Jong NH, Groenhout R, Schoonen R, Hulstijn JH. Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. Appl Psycholinguist. 2015; 36(2):223–43.

48. Fletcher J. The Prosody of Speech: Timing and Rhythm. In: Hardcastle WJ, Laver J, Gibbon FE, editors. The handbook of phonetic sciences. 2nd ed. Hoboken: Wiley-Blackwell; 2010. p. 523–602.

49. Duez D. Silent and non-silent pauses in three speech styles. Lang Speech. 1982; 25(1):11–28.

50. Bybee J. Frequency of Use and the Organization of Language. Oxford: Oxford University Press; 2007.

51. Nespor M, Shukla M, Mehler J. Stress-timed vs. syllable-timed languages. In: Oostendorp M Van, Ewen CJ, Hume E, Rice K, editors. The Blackwell companion to phonology. Malden, MA: John Wiley & Sons; 2011. p. 1–13.

52. Comrie B, editor. The World's Major Languages. 2nd ed. London: Routledge; 2009.

53. Segalowitz N. Cognitive Bases of Second Language Fluency. New York and London: Routledge; 2010.

54. Bilá M, Džambová A. A preliminary study on the function of silent pauses in L1 and L2 speakers of English and German. Brno Stud English. 2011; 37(1):21–39.

55. Cenoz J. Pauses and hesitation phenomena in second language production. In: ITL: Review of Applied Linguistics. 2000. p. 53–69.

56. Boersma P, Weenik D. Praat: doing phonetics by computer [Internet]. 2017. Available from: http://www.praat.org/

57. Deterding D. The North Wind versus a Wolf: short texts for the description and measurement of English pronunciation. J Int Phon Assoc [Internet]. 2006; 36(2):187–96. Available from: https://www.cambridge.org/core/journals/journal-of-the-international-phonetic-association/article/div-classtitlethe-north-wind-versus-a-wolf-short-texts-for-the-description-and-measurement-of-english-pronunciationdiv/984AC3D9FB1F625823E523D2E428B1BE

58. Zellner B. Pauses and the Temporal Structure of Speech. In: Keller E, editor. Fundamentals of speech synthesis and speech recognition. Chichester: John Wiley; 1994. p. 41–62.

59. Shriberg E. To 'errrr' is human: Ecology and acoustics of speech disfluencies. J Int Phon Assoc. 2001; 31(1):153–69.

60. Alkharusi H. Categorical Variables in Regression Analysis: A Comparison of Dummy and Effect Coding. Int J Educ. 2012; 4(2):202.

61. Baayen RH. Analyzing linguistic data. Cambridge: Cambridge University Press; 2008.

62. Box GE, Cox DR. An analysis of transformations revisited, rebutted. J R Stat Soc Ser B. 1964; 26 (2):211–52.

63. Venables WN, Ripley BD. Modern applied statistics with S. 4th ed. New York: Springer; 2002.

64. Baum CF. Stata tip 63: Modeling proportions. Stata J. 2008; 8(2):299–303.

65. Chen K, Cheng Y, Berkout O, Lindhiem O. Analyzing Proportion Scores as Outcomes for Prevention Trials: a Statistical Primer. Prev Sci. 2017; 18(3):312–21. https://doi.org/10.1007/s11121-016-0643-6 PMID: 26960687

66. Lesaffre E, Rizopoulos D, Tsonaka R. The logistic transform for bounded outcome scores. Biostatistics. 2007; 8(1):72–85. https://doi.org/10.1093/biostatistics/kxj034 PMID: 16597671

67. Canty A, Ripley BD. boot: Bootstrap R (S-Plus) Functions. 2017.

68. Schielzeth H, Forstmeier W. Conclusions beyond support: Overconfident estimates in mixed models. Behav Ecol. 2009; 20(2):416–20. https://doi.org/10.1093/beheco/arn145 PMID: 19461866

69. Barr DJ, Levy R, Scheepers C, Tily HJ. Random effects structure for confirmatory hypothesis testing: Keep it maximal. J Mem Lang. 2013; 68(3):255–78.

70. Aarts E, Dolan C V., Verhage M, Van der Sluis S. Multilevel analysis quantifies variation in the experimental effect while optimizing power and preventing false positives. BMC Neurosci. 2015; 16(1):1–15.

71. R Development Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2018. Available from: http://www.r-project.org/

72. RStudioTeam. RStudio: Integrated Development for R. Boston, MA: RStudio, Inc.; 2018.

73. Bates D, Mächler M, Bolker BM, Walker SC. Fitting Linear Mixed-Effects Models Using lme4. J Stat Softw. 2015; 67(1).

74. Field A, Miles J, Field Z. Discovering Statistics Using R. International Statistical Review. Los Angeles: SAGE; 2012.

75. Dobson AJ, Barnett AG. An introduction to generalized linear models. 4th editio. Boca Raton: CRC Press; 2018.

76. Forstmeier W, Schielzeth H. Cryptic multiple hypotheses testing in linear models: overestimated effect sizes and the winner' s curse. Behav Ecol Sociobiol. 2011; 65:47–55. https://doi.org/10.1007/s00265-010-1038-5 PMID: 21297852

77. Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, et al. Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol Evol. 2008; 24(3):127–35.

78. Nakagawa S, Schielzeth H. A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods Ecol Evol. 2013; 4(2):133–42.

79. Bartón K. MuMIn: Multi-Model Inference. R package version 1.42.1 [Internet]. 2018. Available from: https://cran.r-project.org/package=MuMIn

80. Kahng J. Exploring Utterance and Cognitive Fluency of L1 and L2 English Speakers: Temporal Measures and Stimulated Recall. Lang Learn. 2014; 64(4):809–54.

81. Derwing TM, Rossiter MJ, Munro MJ, Thomson RI. Second Language Fluency: Judgements on different tasks. Lang Learn. 2004; 54(4):655–79.

82. Grosjean FE. Temporal variables within and between languages. In: Dechert HW, Raupach M, editors. Towards a Cross-Linguistic Assessment of Speech Production. Bern: Peter Lang; 1980. p. 39–53.

83. Munro MJ, Derwing TM. Modeling perceptions of the accentedness and comprehensibility of L2 speech: the role of speaking rate. Stud Second Lang Acquis. 2001; 23:451–68.

84. Janiszewski C, Wyer RS. Content and process priming: A review. J Consum Psychol [Internet]. 2014; 24(1):96–118. Available from: http://dx.doi.org/10.1016/j.jcps.2013.05.006

85. Gibson S. Reading aloud: A useful learning tool? ELT J. 2008; 62(1):29–36.

Supporting information

Non-native speaker pause patterns closely correspond to those of native speakers at different speech rates

Theresa Matzinger[1,2*], Nikolaus Ritt [1], W. Tecumseh Fitch[2*]

[1]Department of English, University of Vienna, Vienna, Austria
[2]Department of Cognitive Biology, University of Vienna, Vienna, Austria

*** Corresponding authors:**

E-mail: theresa.matzinger@univie.ac.at (TM)

E-mail: tecumseh.fitch@univie.ac.at (WTF)

S1 Results of cross-linguistic studies on pauses

S1 Table. Results of cross-linguistic studies suggesting that the numbers and durations of pauses in different languages are similar.

Characteristics considered	Results	References
Number of pauses	English ≈ French ≈ Spanish	(1)
	French ≈ English	(2)
	English ≈ Croatian	(3)
	adjustments to different speech tempi: Dutch ≈ English ≈ French ≈ Italian ≈ Romanian ≈ Spanish	(4)
Duration of pauses	English ≈ French ≈ Spanish	(1)
	English ≈ Croatian	(3)
	English ≈ Chinese	(5)
	durational adjustments to different speech tempi: Dutch ≈ English ≈ French ≈ Italian ≈ Romanian ≈ Spanish	(4)

S2 Results of cross-linguistic studies on pauses

S2 Table. Results of cross-linguistic studies suggesting that the numbers and durations of pauses in different languages are different.

Characteristics considered	Results	References

Number of pauses	English > French	(6,7)
	Spanish > English	(8)
	English > Turkish	(9)
	French > German	(10)
Duration of pauses	French > English	(2,6)
	Russian > English	(11)
	French > German	(10)
	Spanish > English ≈ French ≈ German > Italian	(12)
	durational change in older age: German > French	(13)

S3 Results of studies on pauses in L2 speech

S3 Table. Results of studies on the numbers and durations of pauses during L2 speech. Results in the table concern comparisons between pauses in speakers' L2s and these speakers' L1s (as opposed to L1 speakers of the target L2).

Characteristics considered	**Results**	**L1**	**L2**	**References**
Number of pauses/Pause rate	L2 = L1	Russian, Ukrainian, Mandarin	English	(14)
		Japanese	English	(15)
	L2 > L1	Russian	English (intermediate and high proficiency)	(11)
		Thai	English	(16)
		French	English	(17)
		German	French	(2,10)
		French	German	(2,10)
		English, Turkish	Dutch	(9)
Number of pauses between utterances	L2 = L1	English, Turkish	Dutch	(18)
Number of pauses within utterances	L2 > L1	English, Turkish	Dutch	(18)
		German	French	(10)
		French	German	(10)
		"a range of different L1 backgrounds"	English	(19)
Duration of pauses	L2 = L1	Russian, Ukrainian, Mandarin	English	(14)
		Japanese	English	(15)
		Russian	English (intermediate proficiency)	(11)
		French	English	(17)
	L2 > L1	German	French	(2)
		Turkish, English	Dutch	(18)
		German	English, French	(20)
		English, Turkish	Dutch	(9)
	L2 < L1	German	French	(10)
		French	German	(2,10)
		English	Russian (high proficiency)	(11)

Duration of pauses between utterances	L2 = L1	English, Turkish	Dutch	(18)
Duration of pauses within utterances	L2 > L1	English, Turkish	Dutch	(18)

S4 Results of studies on pauses in L2 speech

S4 Table. Results of studies on the numbers and durations of pauses during L2 speech. Results in the table concern comparisons between pauses in speakers' L2s and L1 speakers of the target L2.

Characteristics considered	Results	L1	L2	References
Number of pauses/Pause rate	L2 = L1	Russian	English (high proficiency)	(11)
	L2 > L1	Russian	English (intermediate proficiency)	(11)
		Korean	English	(21)
Duration of pauses	L2 = L1	Russian	English (high proficiency)	(11)
	L2 < L1	Russian	English (intermediate proficiency)	(11)

S5 Appendix. Post-experiment questionnaire.

Questionnaire for participants

ID: ______________________________

Age: ______________________________ Date: ______________________________

Gender: ______________________________ Nationality: ______________________________

PLEASE ANSWER THE FOLLOWING QUESTIONS:

Part I: Your language(s)

1) What is your first language (please include not only the standard but also non-standard variety (i.e. your "dialect")?

2) Did you grow up bilingually? If so, what are your first languages?

3) Which other languages do you speak (please specify estimated proficiency level)?

Part II: Taking a closer look at your English (only for participants whose first language is not English)

1) At what age did you start learning English?

2) Which variety of English do you aim to speak?

3) Have you ever taken English pronunciation classes? Please specify (For which variety? For how long? etc.)

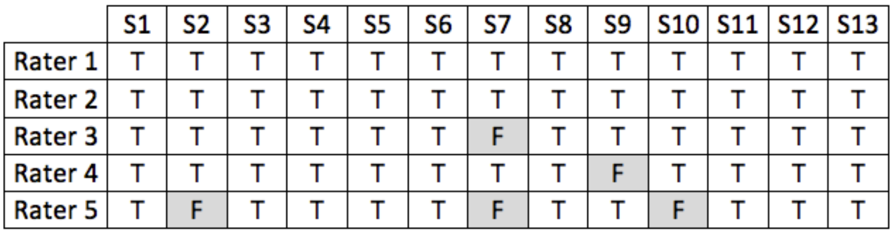
4) Do you aspire to have native-like pronunciation?

5) Which other influences might have shaped the way you speak English (i.e. for example having lived in an English-speaking country for a period of time, media exposure, etc.)?

S6 Native language recognition ratings.

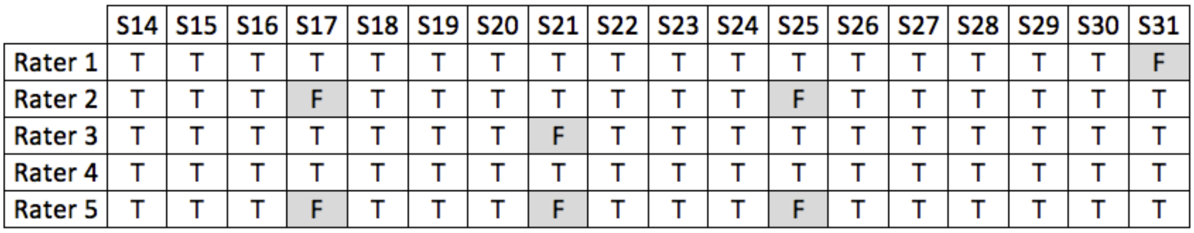
S6 Table. Results of a native language recognition test for our pool of speech samples of English, German and Serbo-Croatian native speakers (S1-S41) reading an English text. T = native English raters identified the speakers' nativeness status correctly. F (gray) = native English raters misclassified the speakers' nativeness status. Four German native speakers were misclassified as being English native speakers by one or two of the raters each, but correctly identified as being non- native by the other raters. Four English native speakers were misclassified as being non-native speakers of English by one or two of the raters each, but correctly identified as native speakers by the other raters.

English native speakers



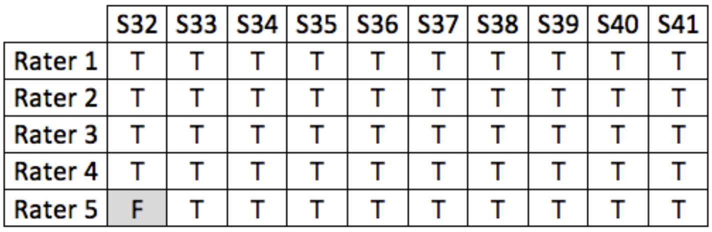
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13
Rater 1	T	T	T	T	T	T	T	T	T	T	T	T	T
Rater 2	T	T	T	T	T	T	T	T	T	T	T	T	T
Rater 3	T	T	T	T	T	T	F	T	T	T	T	T	T
Rater 4	T	T	T	T	T	T	T	T	F	T	T	T	T
Rater 5	T	F	T	T	T	T	F	T	T	F	T	T	T

German native speakers



	S14	S15	S16	S17	S18	S19	S20	S21	S22	S23	S24	S25	S26	S27	S28	S29	S30	S31
Rater 1	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	F
Rater 2	T	T	T	F	T	T	T	T	T	T	T	F	T	T	T	T	T	T
Rater 3	T	T	T	T	T	T	T	F	T	T	T	T	T	T	T	T	T	T
Rater 4	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
Rater 5	T	T	T	F	T	T	T	F	T	T	T	F	T	T	T	T	T	T

Serbo-Croatian native speakers



	S32	S33	S34	S35	S36	S37	S38	S39	S40	S41
Rater 1	T	T	T	T	T	T	T	T	T	T
Rater 2	T	T	T	T	T	T	T	T	T	T
Rater 3	T	T	T	T	T	T	T	T	T	T
Rater 4	T	T	T	T	T	T	T	T	T	T
Rater 5	F	T	T	T	T	T	T	T	T	T

S7 Appendix. Annotated text *The boy who cried wolf.*

For determining the positions of pauses, we used the following classification:

[MA] pauses occurring at punctuation marks
[UM] pauses occurring at unmarked clause or phrase boundaries

Pauses at other positions in the text (e.g. between "There" and "was", or between "was" and "once"), were classified as "other" and are not indicated in the text below. In total, 76.29 % of all pauses made occurred at punctuation marks, 15.13 % occurred at unmarked clause or phrase boundaries, and 8.58 % occurred at other positions in the text.

<u>The Boy who Cried Wolf</u>

There was once a poor shepherd boy **[UM]** who used to watch his flocks **[UM]** in the fields **[UM]** next to a dark forest **[UM]** near the foot of a mountain. **[MA]** One hot afternoon, **[MA]** he thought up a good plan **[UM]** to get some company for himself **[UM]** and also have a little fun. **[MA]** Raising his fist in the air, **[MA]** he ran down to the village **[UM]** shouting ‘ **[MA]** Wolf, **[MA]** Wolf.’ **[MA]** As soon as they heard him, **[MA]** the villagers all rushed from their homes, **[MA]** full of concern for his safety, **[MA]** and two of his cousins even stayed with him for a short while. **[MA]** This gave the boy so much pleasure **[UM]** that a few days later **[UM]** he tried exactly the same trick again, **[MA]** and once more he was successful. **[MA]** However, **[MA]** not long after, **[MA]** a wolf **[UM]** that had just escaped from the zoo **[UM]** was looking for a change **[UM]** from its usual diet **[UM]** of chicken and duck. **[MA]** So, **[MA]** overcoming its fear of being shot, **[MA]** it actually did come out from the forest **[UM]** and began to threaten the sheep. **[MA]** Racing down to the village, **[MA]** the boy of course cried out even louder **[UM]** than before. **[MA]** Unfortunately, **[MA]** as all the villagers were convinced **[UM]** that he was trying to fool them a third time, **[MA]** they told him, ‘ **[MA]** Go away **[UM]** and don’t bother us again.’ **[MA]** And so the wolf had a feast.

S8 Results of the random effects of the linear mixed model exploring the effects of reading tempo and nativeness on the total reading time.

S8 Table. Estimated variance components and standard deviations for the random intercept of participant of the full model exploring the effects of reading tempo, and nativeness on the total reading time.

Random effect	**Term**	**Variance**	**Standard deviation**
Participant	Intercept	0.01	0.08
Residual		0.01	0.11

S9 Results of the random effects of the linear mixed model exploring the effects of reading tempo and native language on pause-to-utterance ratio.

S9 Table. Estimated variance components and standard deviations for the random intercept of participant of the full model exploring the effects of reading tempo and nativeness on pause-to-utterance ratio.

Random effect	**Term**	**Variance**	**Standard deviation**
Participant	Intercept	0.07	0.27
Residual		0.07	0.26

S10 Results of the random effects of the linear mixed model exploring the effects of reading tempo and native language on the duration of individual pauses.

S10 Table. Estimated variance components and standard deviations for the random intercept of participant of the full model exploring the effects of reading tempo, and nativeness on the duration of individual pauses.

Random effect	**Term**	**Variance**	**Standard deviation**

Participant	Intercept	0.04	0.20
Residual		0.03	0.17

S11 Results of the random effects of the logistic regression model exploring the effects of reading tempo, nativeness and in-text position on the occurrence frequency of pauses.

S11 Table. Estimated variance components and standard deviations for the random intercept of participant of the full model exploring the effects of reading tempo, nativeness and in-text position on the occurrence frequency of pauses.

Random effect	Term	Variance	Standard deviation
Participant	Intercept	0.32	0.56

References

1. Huensch A, Tracy-Ventura N. Understanding second language fluency behavior: the effects of individual differences in first language fluency, cross-linguistic differences, and proficiency over time. Appl Psycholinguist. 2016;1–31.

2. Trouvain J, Fauth C, Möbius B. Breath and Non-breath Pauses in Fluent and Disfluent Phases of German and French L1 and L2 Read Speech. In: Speech Prosody (SP8) [Internet]. 2016. p. 31–5. Available from: http://www.ifcasl.org/docs/Trouvain_Fauth_Moebius_2016.pdf

3. Smiljanic R, Bradlow AR. Production and perception of clear speech in Croatian and English. J Acoust Soc Am. 2005;118:1677–88.

4. Demol M, Verhelst W, Verhoeve P. The duration of speech pauses in a multilingual environment. Proc Annu Conf Int Speech Commun Assoc INTERSPEECH. 2007;1(1):117–20.

5. Yang L. Duration and pauses as boundary-markers in speech: a cross-linguistic study. In: Proceedings of Interspeech 2007 [Internet]. 2007. p. 458–61. Available from: http://www.speech.kth.se/prod/publications/files/100444.pdf

6. Grosjean FE, Deschamps A. Analyse contrastive des variables temporelles de l'anglais et du francais: vitesse de parole et variables composantes, phénomènes d'hésitation. Phonetica. 1975;31:144–84.

7. Holmes VM. A crosslinguistic comparison of the production of utterances in discourse. Cognition. 1995;54:169–207.

8. Johnson TH, O'Connell DC, Sabin EJ. Temporal analysis of English and Spanish narratives. Bull Psychon Soc. 1979;13(6):347–50.

9. De Jong NH, Groenhout R, Schoonen R, Hulstijn JH. Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. Appl Linguist. 2013;36:1–21.

10. Raupach M. Temporal variables in first and second language speech and perception of fluency. In: Dechert HW, Raupach M, editors. Temporal variables in speech: Studies in honour of Frieda Goldman-Eisler [Internet]. The Hague: Mouton; 1980. p. 263–70. Available from: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0405.pdf

11. Riazantseva A. Second Language Proficience and Pausing: A Study of Russian Speakers of English. Stud Second Lang Acquis [Internet]. 2001;23(4):497–526. Available from: http://www.journals.cambridge.org/abstract_S027226310100403X

12. Campione E, Véronis J. A large-scale multilingual study of pause duration. Speech Prosody 2002 Proc 1st Int Conf Speech Prosody [Internet]. 2002;199–202. Available from: http://www.isca-speech.org/archive/sp2002/sp02_199.html

13. Gerstenberg A, Fuchs S, Kairet JM, Frankenberg C, Schröder J. A cross-linguistic , longitudinal case study of pauses and interpausal units in spontaneous speech corpora of older speakers of German and French. In: Proceedings of the 9th International Conference on Speech Prosody. 2018. p. 1–5.

14. Derwing TM, Munro MJ, Thomson RI, Rossiter MJ. The relationship between L1 fluency and L2 fluency development. Stud Second Lang Acquis. 2009;31(4):533–57.

15. Rose R. Temporal variables in first and second language speech and perception of fluency. ICPhS 2015 Proc 18th Int Congr Phonetic Sci [Internet]. 2015;1–5. Available from: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0405.pdf

16. Isarankura S. Variability of Pause Patterns in English Read Speech of Thai EFL Learners. J Educ Soc Res [Internet]. 2013;3(7):346–54. Available from: http://www.mcser.org/journal/index.php/jesr/article/view/971

17. Deschamps A. The syntactical distribution of pauses in English spoken as a second language by French students. In: Dechert HW, Raupach M, editors. Temporal Variables in Speech: Studies in honour of Frieda Goldman-Eisler. The Hague: Mouton; 1980. p. 255–62.

18. De Jong NH. Predicting pauses in L1 and L2 speech: the effects of utterance boundaries and word frequency. Int Rev Appl Linguist Lang Teach. 2016;54(2):113–32.

19. Tavakoli P. Pausing patterns: Differences between L2 learners and native speakers. ELT J. 2011;65(1):71–9.

20. Kolly M-J, Leemann A, Boula P, Mareüil D, Dellwo V. Speaker-idiosyncrasy in pausing behavior: evidence from a cross-linguistic study. Proc Int Congr Phonetic Sci 2015 Glas. 2015;1–5.

21. Trofimovich P, Baker W. Learning Second Language Suprasegmentals: Effect of L2 Experience on Prosody and Fluency Characteristics of L2 Speech. Stud Second Lang Acquis. 2006;28(1):1–30.

CHAPTER 5

Phonotactically probable word shapes represent attractors in the evolution of sound patterns

This chapter is under review in *Cognitive Linguistics*.

Citation:

Matzinger T & Ritt N. Phonotactically probable word shapes represent attractors in the evolution of sound patterns (*under review*)

Phonotactically probable word shapes represent attractors in the cultural evolution of sound patterns

Keywords

frequency effects; sound change; Middle English; Open Syllable Lengthening; cultural evolution

Abstract

Words are processed more easily when they have canonical phonotactic shapes, i.e. shapes that are frequent both in the lexicon and in word usage. We hypothesized that words with such shapes have a selective advantage in cultural evolution and favour sound changes that (re-)produce words with the same shapes. We tested this hypothesis in a quantitative corpus study on the Middle English sound change known as Open Syllable Lengthening (OSL). OSL lengthened vowels in disyllables such as ME /ma.kə/ *make*, but more or less only when they became monosyllabic and when their vowels were non-high. We predicted that word shapes produced by this implementation pattern should correspond to the shapes that were most common among morphologically simple monosyllables and disyllables at the time when OSL occurred. Our results largely confirmed this prediction: monosyllables produced by OSL did indeed conform to the shapes that were most frequent among already existing monosyllables. Similarly, the failure of OSL to affect disyllables (such as *body*) prevented them from assuming shapes that were far more typical of morphologically complex word forms than of simple ones. This suggests that the actuation and implementation of sound changes may be more sensitive to lexical probabilities than hitherto suspected.

1 Introduction

This paper deals with the hypothesis that the cultural evolution of sound patterns is constrained by a preference for probable word shapes because their relative frequency makes them predictable and easier to identify and process. This implies that word shapes that already are in the majority, should be selected for and get even more frequent, whereas word shapes that are in the minority should be selected against and get less frequent.

1.1 The role of majority patterns in the cultural evolution of sound patterns

Our starting point is the observation that speakers are sensitive to the probability of phonotactic patterns in the lexical inventory (Wedel 2006; Blevins 2009; Mailhammer, Kruger, and Makiyama 2015; Divjak 2019a, 2019b), and exploit it in the perception and the production of speech (Gibson et al. 2019). Word forms whose shapes are probable – i.e. represent majority patterns – are identified more easily (Kelley and Tucker 2017), learnt and memorized more easily (Storkel 2001), and produced more accurately (Goldrick and Larson 2008) than less frequent and less probable forms. Since word forms that are more easily recognized, memorized, and used, will also be transmitted more easily, word forms with more probable shapes ought to be historically more stable than word forms with rare or exceptional shapes. Thus, word form shapes that happen to be probable in a specific language will select for word form variants that conform to them, and function as attractors in phonological evolution (Blevins 2006, 2009). This implies that they ought to favour sound changes that (re-)produce them and thus stabilize or even bolster their own majority. This will further increase the advantages they afford in terms of the recognition, processing, and use of words. Crucially, this ought to be the case even if preferences for majority patterns are weak, since language history reflects the outcome of massively parallel and iterated transmission processes, which are known to amplify even very weak biases (Fehér, Wonnacott, and Smith 2016; Kirby, Griffiths, and Smith 2014; Smith et al. 2017; Reali and Griffiths 2009).

To test the hypothesis that probable word shapes represent attractors in the diachronic development of sound patterns, we investigate whether it predicts the lexical implementation of a specific Middle English sound change, namely Open Syllable Lengthening (henceforth OSL; Ritt 1994; Minkova 1982; Luick 1964). OSL was a sound change which lengthened vowels in open syllables and which, as a result, produced forms such as Late Middle English /maːk/ *make* or /beː.vər/ *beaver* from earlier /ma.kə/ and /be.vər/. We show that the words in which the change was regularly implemented did indeed get shapes that represented majority patterns. However, in words where it would have produced minority patterns, OSL was implemented only sporadically. Therefore, our paper provides evidence for the theory that the actuation and implementation of sound changes may be conditioned by the statistical distribution of word-shapes in the lexicon. Additionally, it deepens our understanding of OSL, a sound change with a complex implementation pattern.

1.2 The test case: Middle English Open Syllable Lengthening

We investigated the lexical implementation of Middle English Open Syllable Lengthening and asked if it reflected a preference for probable sound patterns. We chose this particular sound change because it has an interesting implementation pattern, which has been well described but not fully explained. Thus, choosing OSL as a test case for our study, we not only test a general hypothesis on the cultural evolution of sound patterns, but also contribute to explaining a specific and widely discussed sound change in the history of English (Ritt 1994; Minkova 1982; Mailhammer, Kruger & Makiyama 2015, Minkova & Lefkovitz 2020).

OSL occurred between the 13th and 14th centuries and lengthened short vowels in stressed open syllables in disyllabic words, such as *make* (/ma.kə/ > /maːk/), *hope* (/hɔ.pə/ > /hɔːp/), *name* (/na.mə/ > /naːm/), or *beaver* (/be.vər/ > /beː.vər/) (Ritt 1994; Minkova 1982; Mailhammer, Kruger & Makiyama 2015, Minkova & Lefkovitz 2020). However, there were crucial restrictions on its implementation. First, OSL did not regularly affect words with high vowels, such as *sin* (/si.nə/ > */siːn/)[1]. Second, the implementation of OSL depended on the phonotactic structure of the words it affected: it affected those words consistently that were originally disyllabic, but became monosyllabic through the loss of their final syllable (a change known as schwa loss, as in /ma.kə/ > /maːk/), which occurred roughly at the same period as OSL (Tab. 1; Minkova 1991, Minkova & Lefkovitz 2020, Minkova forthc.). In contrast, it affected words that remained disyllabic only rarely. Most of the few stable disyllables that were lengthened had sonorants in their second syllable (e.g. *beaver*, *bacon*), and a single one an obstruent (*naked*). Other stable disyllables – like *body* or *many* – were never lengthened (Tab. 1). Thus, the only items in which the change was implemented nearly categorically were disyllables that – like *make*, *name* or *hope* – had non-high vowels and became monosyllabic due to final schwa loss.

[1] Some rare exceptions that were lengthened despite their high vowels are *door* (/du.rə/ > /doːr/), *beetle* (/bi.tʊl/ > /beː.təl/), or *evil* (/y.vəl/ > /eː.vəl/).

Table 1. Implementation patterns of OSL, depending on the phonotactic structure of words (C = consonant, V = vowel, . = syllable boundary, ə = schwa, R = sonorant, T = obstruent; reproduced from Minkova and Lefkowitz 2020)

Phonotactic structure	Examples	Percentage of implementation
CV.Cə	*make, hope, name*	95.1%
CV.CVR	*beaver*, *hammer*	22.9%
CV.CVT	*habit*, *naked*	2.5%
CV.CV	*body*, *many*	0%

A question that has intrigued historical linguists about this implementation pattern is why OSL appears to have been implemented primarily in words where the phonological conditions that motivated the lengthening in the first place, i.e. the open syllables, were lost. After all, when OSL was completed, the vowels in lengthened *make* /maːk/, *name* /naːm/ or *hope* /hɔːp/ were no longer in open syllables at all.[2] Thus, it seems as if OSL affected vowels only in such open syllables that were in the process of becoming closed through the loss of final schwas. The conclusion that has been drawn from this perfectly adequate observation is that the lengthening must have been compensatory, i.e. it made up for the weight loss induced by schwa loss (Minkova 1982; Bermudez-Otero 1998; Minkova and Lefkowitz 2020). This hypothesis receives support from the fact that stable disyllables in which schwa loss was at least optional (e.g. *beaver* could be realized as /beː.vr̩/ or /beː.vər/) were also occasionally lengthened (Tab. 1). Although this account is descriptively adequate, it still raises the question why compensation should have occurred, because the existence of forms such as *man* /man/ and *god* /gɔd/ suggests that realizations with short vowels such as */mak/, */nam/ and */hop/ would have been just as viable as the long realizations /maːk/, /naːm/ and /hɔːp/.

2 Hypotheses and predictions

2.1 General hypotheses

We explore the possibility that compensatory lengthening occurred in words of the *make*-type but not (or much less so) in the *beaver*/*habit*/*body*-type for statistical reasons. Specifically, we hypothesize that phonotactic patterns similar to the ones of lengthened /maːk/ should have been more frequent among already existing morphologically simple monosyllabic words than patterns similar to the ones of hypothetical unlengthened */mak/. If that was the case, the lengthened variants of *make*-type items could have been selected and historically stabilized for that very reason. While unlengthened realizations like */mak/ would not have been strictly speaking unviable, the higher probability of their lengthened competitors would have made lengthened forms easier to identify and process, so that they were ultimately selected. Conversely, the opposite should have applied to OSL outputs that remained disyllabic. In their case, the shapes of lengthened items such as hypothetical *body* */bɔː.dij/ or *habit* */haː.bit/ should have been less representative of morphologically simple disyllables than unlengthened variants such as the attested /bɔ.dij/ and /ha.bit/.

If both of these hypotheses turn out to be true, then the implementation of OSL would have increased the frequency of word forms with prototypical shapes. This would have facilitated the identification, processing and use not only of the OSL reflexes themselves, but also of all

[2] Note that vowels in words like *man* or *god*, which had never been in an open syllable, had not lengthened. So it is indeed strange that open syllables should have motived vowel lengthening, but only if they were becoming closed.

other words that shared the same phonotactic shapes, so that the overall efficiency of the lexicon would have risen. Thus, the implementation of OSL could ultimately be explained as being conditioned by (a) the frequency distribution of word form shapes in the lexicon, and (b) a universal preference for that distribution to be skewed in favour of prototypical shapes.

In order to find out if our hypotheses were plausible, we categorized and counted the phonotactic shapes of monosyllabic and disyllabic wordforms attested in Early Middle English. We analysed if the frequencies of existing Early Middle English monosyllables with shapes similar to OSL outputs were higher than those of monosyllables with shapes different from OSL outputs. Similarly, we compared the frequencies of various phonotactic shapes of existing Early Middle English disyllables. For these tasks, several theoretical decisions had to be taken, which will be described in the following sections.

2.2 Theoretical considerations and predictions for Early Middle English monosyllables

An important theoretical question was which phonotactic shapes of existing Early Middle English word forms (Tab. 2b) should be counted as being ‘similar’ or ‘dissimilar’ to the ones of lengthened OSL outputs (Tab. 2a). The patterns of already existing Early Middle English word forms can be compared with those of OSL outputs in two ways: either in terms of their vowel length, i.e. long (VV) vs. short (V), or more broadly in terms of their syllable weight, i.e. heavy (VVC, VCC, VVCC) vs. light (VC; Tab. 2). The chosen level of comparison has consequences for the predictions that follow.

Table 2. Phonotactic structure of a. monosyllabic OSL outputs, and b. already existing Early Middle English word forms that act as candidates for comparison.[3] Final consonants are counted as extrametrical.

	a. Lengthened OSL outputs	**b. Existing Early Middle English monosyllabic word forms**			
	***make*-type**	***god*-type**	***mood*-type**	***land*-type**	
Phonotactic shape	CVVC	CVC	CVVC	CVCC	CVVCC
Pronunciation	/ma:k/	/gɔd/	/moːd/	/land/	/gruːnd/
Vowel length	long	short	long	irrelevant for our classification	
Syllable weigth	heavy	light	heavy	heavy	

Regarding syllable weight, and considering final consonants as extrametrical, *god*-type items count as light, and *mood*-type and *land*-type items as heavy (Tab. 2b). OSL outputs, i.e. *make*-type items, count as heavy (Tab. 2a) and would thus be ‘similar’ in their syllable weight to all other heavy monosyllables. Following our general hypothesis, we therefore predict that at the time when OSL and schwa loss set in, heavy Early Middle English monosyllables (*mood*-type and *land*-type items) were in the majority compared to light ones (*god*-type items; see Fig. 1, Prediction 1). If this were the case, these frequency patterns would select for heavy and thus lengthened OSL outputs, and against unlengthened competitor variants such as the hypothetical light */mak/.

[3] Note that *CV was ungrammatical for major class items already in Middle English.

A stricter and more specific comparison would focus on vowel length as the only relevant variable, and would disregard items that are heavy just because they end in consonant clusters. Thus, this stricter comparison only considers *god*-type and *mood*-type items and discards *land*-type items. On this stricter criterion, our hypothesis would be corroborated if words with long vowels, i.e. phonotactic shapes like *mood* were more frequent in Early Middle English than words with short vowels like *god* (see Fig. 1, Prediction 2).

Finally, an even stricter comparison would also take vowel height into account. Recall that high vowels were affected by OSL only very sporadically. Thus, our hypothesis predicts that words from the *mood*-type should only be more frequent than words from the *god*-type, when these words had mid or low vowels such as /a/, /e/, /ɔ/ or /o/. In contrast, among words with high vowels such as /i/ or /u/, long vowels should not be more frequent than short ones (see Fig. 1, Prediction 3).

In our study we applied all three types of measure. In summary, the weakest version of our hypothesis was that the majority of monosyllables in our data should be heavy, the stronger version was that we should find more monosyllables with long than short vowels, and the strongest version was that we should find more items with long vowels only among words with non-high vowels, but not among words with high ones (Fig. 1).

Figure 1. Summary of predictions for monosyllables.

Predictions for Early Middle English monosyllables

① Among monosyllables, the majority of items was heavy, i.e. belonged to the *mood*-type or *land*-type.

② Among monosyllables ending in single consonants, the majority of items had long vowels or diphthongs, i.e. belonged to the *mood*-type.

③ Prediction 2 was only true for items with mid or low vowels, but not for items with high vowels.

2.3 Theoretical considerations and predictions for Early Middle English disyllables

In the case of disyllables, we again had to decide if we should focus on the syllable weight or just on the vowel length of their first syllables, when we counted them as being 'similar' to disyllabic OSL reflexes whose first syllable got lengthened (*beaver*-type items; Tab. 3a) or remained short (*habit*-type items; Tab. 3a). Once again, we worked with both measures: when focusing on syllable weight, we included heavy but closed first syllables (*finger*-type items; Tab. 3b) in the comparison. When exclusively focussing on open first syllables, we merely counted how many of them had long vowels or diphthongs (*bailiff*-type items) and how many of them short ones (*mother*-type items; Tab. 3b).

Additionally, however, we also had to take the morphological structure of the disyllables into account. Many disyllables were in fact morphologically complex, such as *doom-es* 'doom.GEN', *drench-es* ‘drink.PL’, or *sorh-en* ‘sorrow.PL’. In this respect, they were unlike the lengthened or unlengthened reflexes of OSL inputs, which – like *beaver* – were all simple. Thus, the question was not only how probable the shapes of OSL outputs and their unlengthened variants were as exponents of words, but how probable each of the two were as exponents of morphologically simple words. Morphologically simple word forms with shapes like complex ones invite unwarranted decomposition, and delay identification and processing (Post et al. 2008). This affects not only the ease with which word forms are acquired but also their historical stability, which has been shown in various studies in the area of morphonotactics (Dressler and Dziubalska-Kołaczyk 2006; Calderone et al. 2014; Korecky-Kröll et al. 2014; Ritt and Kaźmierski 2015; Baumann and Kaźmierski 2018; Baumann, Prömer, and Ritt 2019). Therefore, morphologically simple disyllabic OSL outputs that were shaped like morphologically complex disyllables would not be identified and processed more easily because of that similarity. On the contrary, their similarity with complex disyllables would make their identification and processing more difficult and would therefore be select against. Thus, in general, we expected morphologically simple disyllabic OSL outputs to have shapes similar to existing Early Middle English morphologically simple disyllables and dissimilar to morphologically complex ones.

Table 3. Phonotactic shape, syllable structure and syllable weight of the first syllable (σ_1) as well as examples of a. disyllabic reflexes of OSL inputs and b. already existing Early Middle English disyllables that act as candidates for comparison.

	a. Disyllabic reflexes of OSL inputs		**b. Existing Early Middle English disyllabic word forms**		
	lengthened (rare)	**unlengthened (frequent)**			
	***beaver*-type**	***habit*-type**	***mother*-type**	***bailiff*-type**	***finger*-type**
Phonotactic shape of σ_1	CVV	CV	CV	CVV	CVC
Morphologically simple examples	bea.ver	ha.bit	mo.ther bi.shop	bai.liff	fin.ger mer.cy an.gel
Morphologically complex examples	irrelevant for our study	irrelevant for our study	ta.l-es no.m-en	doo.m-es	dren.ch-es, sor.h-en
Syllable structure	open	open	open	open	closed
Vowel length	long	short	short	long	irrelevant for our classification
Syllable weight	heavy	light	light	heavy	heavy

In sum, when we categorized the disyllables that existed at the time when OSL set in, we therefore had to distinguish not only between their phonotactic shapes, i.e. between words with light and therefore short open first syllables (such as *mo.ther*), words with heavy open long first syllables (such as *bai.liff*), and words with closed (and therefore heavy) first syllables (such as *fin.ger*), but also – on the morphotactic level – between simple words like *mo.ther* or *bi.shop*, and complex words like *doo.m-es*, or *dren.ch-es*.

Recall that only a minority of morphologically simple OSL inputs that remained disyllabic was lengthened (such as *bea.ver*) and the majority remained short (such as *ha.bit*). Therefore, we predicted that (a) the majority of Early Middle English disyllables with light, i.e. open short first syllables would have been simple (such as *mo.ther*; see Fig. 2, Predictions 1 & 3) and that (b) that the majority of disyllables with heavy first syllables (or with long open ones, depending on the chosen measure) would have been complex (such as *doo.m-es* or *dren.ch-es*; see Fig. 2, Predictions 2 & 4). If this prediction turned out to be true, this would mean that morphological simplicity of disyllables was indicated by light first syllables and morphological complexity by heavy or long first syllables. This would support that the reason why *habit-type* items were not affected by OSL was that OSL in those items would have made perception less intuitive.

Another aspect that we had to consider in the case of disyllables was the directionality of correlations. On the one hand one can ask how likely it is that a specific phonotactic shape stands for either a simple or a complex word (Tab. 4a; and as outlined in the previous paragraph), and on the other hand one can ask how likely it is for a simple or complex item to be represented by a specific phonotactic shape (Tab. 4b).

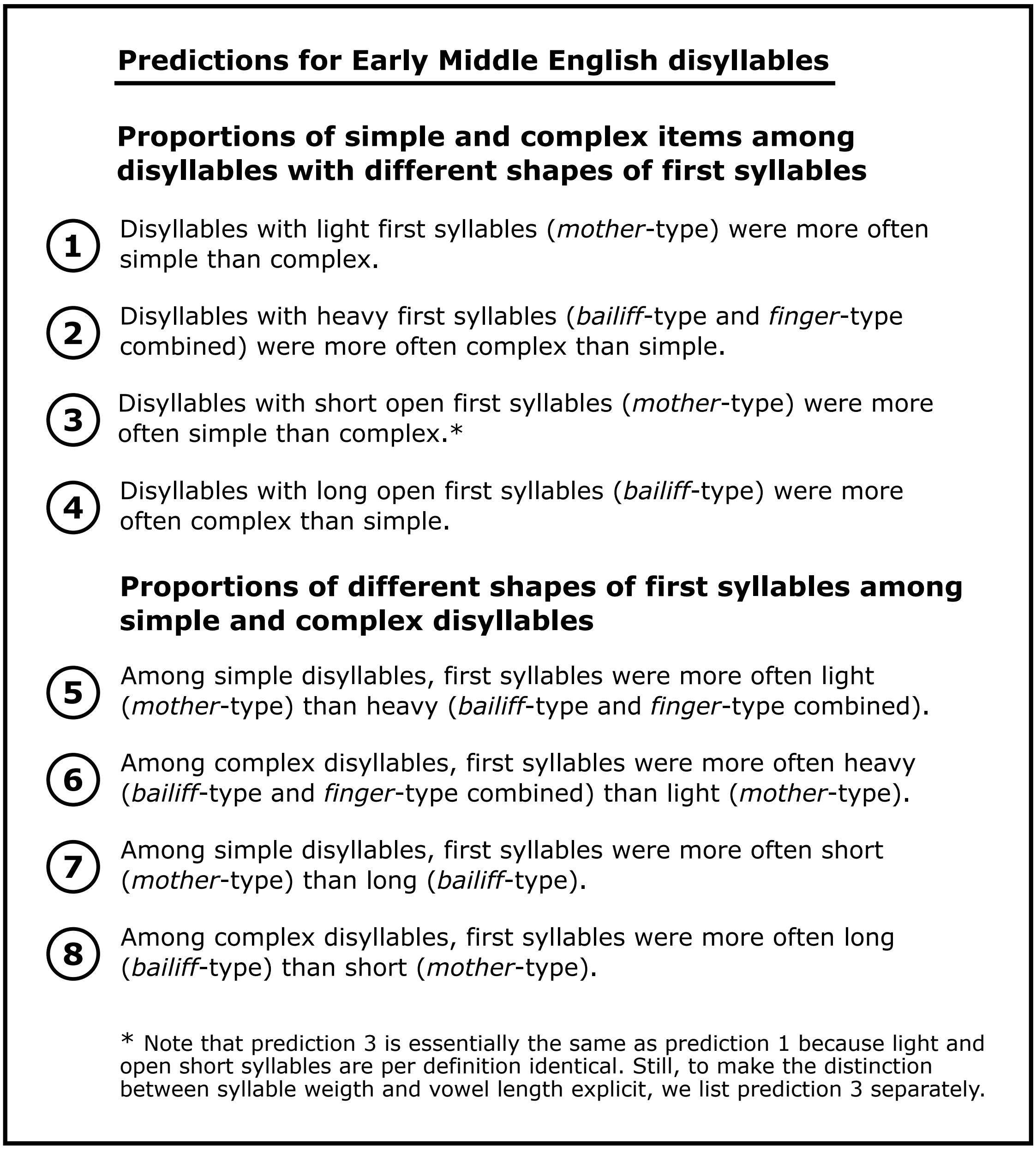
Table 4. Directionalities of correlations

	What language users encounter or expect	→	**what language users can infer from that**
a.	light first syllable (e.g. **mo**.ther)	→	high probability that the item is simple
	heavy first syllable (e.g. **doo**.mes)	→	high probability that the item is complex
b.	simple item (e.g. {mother})	→	high probability that the first syllable is light
	complex item (e.g. {doom}+{es})	→	high probability that the first syllable is heavy

Both directionalities affect the identification and the processing of words and their morphotactic structures. For the relationship in Tab. 4a, this is obvious: if language users know that most items with light first syllables stand for simple words, this will help them to identify such shapes as simple words when they hear them. However, also the correlations in Tab. 4b are helpful for language processing. This is because perception is influenced by expectations (McClelland and Elman 1986; Cole, Mo, and Hasegawa-Johnson 2010; de Lange, Heilbron, and Kok 2018). For example, if context makes listeners expect a morphologically simple word (e.g. the base form of a noun such as *bishop*), and if they know that simple words are more likely to have light than heavy first syllables, it will be easier for them to perceive this word when it has indeed a light first syllable.

Therefore, our hypothesis will be corroborated most strongly, if all directionalities in Tab. 4 are supported by our data. This means that the majority of items with a light first syllable should be simple and the majority of items with a heavy first syllable should be complex (Tab. 4a; Fig. 2, Predictions 1, 2, 3 & 4). Similarly, the majority of simple items should have a light and short first syllable and the majority of complex items should have heavy (or long) first syllable (Tab. 4b; Fig. 2, Predictions 5, 6, 7 & 8).

Figure 2. Summary of predictions for disyllables.



Predictions for Early Middle English disyllables

Proportions of simple and complex items among disyllables with different shapes of first syllables

① Disyllables with light first syllables (*mother*-type) were more often simple than complex.

② Disyllables with heavy first syllables (*bailiff*-type and *finger*-type combined) were more often complex than simple.

③ Disyllables with short open first syllables (*mother*-type) were more often simple than complex.*

④ Disyllables with long open first syllables (*bailiff*-type) were more often complex than simple.

Proportions of different shapes of first syllables among simple and complex disyllables

⑤ Among simple disyllables, first syllables were more often light (*mother*-type) than heavy (*bailiff*-type and *finger*-type combined).

⑥ Among complex disyllables, first syllables were more often heavy (*bailiff*-type and *finger*-type combined) than light (*mother*-type).

⑦ Among simple disyllables, first syllables were more often short (*mother*-type) than long (*bailiff*-type).

⑧ Among complex disyllables, first syllables were more often long (*bailiff*-type) than short (*mother*-type).

* Note that prediction 3 is essentially the same as prediction 1 because light and open short syllables are per definition identical. Still, to make the distinction between syllable weigth and vowel length explicit, we list prediction 3 separately.

2.4 Type vs. token frequencies

When counting frequencies of words with specific phonotactic patterns and morphological structures, we took both type and token frequencies into account. This is because the production, perception and processing of sound shapes may be influenced both by the number of types, i.e. different word forms that have these sound shapes, and the number of tokens, i.e. utterances in which they occur (Berg 2014). Type frequencies have been shown to be better predictors for phonological and morphological pattern learning than token frequencies (Richtsmeier 2011; Pierrehumbert 2016; Baumann, Prömer, and Ritt 2019; Bybee 1995 – but see Baumann and Kaźmierski 2018 for counterevidence). We therefore predicted that types would show more distinctive majority patterns than tokens.

3 Methods

3.1 Data collection

To test if the frequency patterns of phonotactic shapes and morphological structures in Early Middle English word forms follow our predictions, we used Early Middle English corpus data from the LAEME corpus (The Linguistic Atlas of Early Middle English; Laing 2013). We chose LAEME because it covers the period in which schwa loss and OSL began to unfold (1150-1325), and also because it is lemmatized and grammatically tagged at a high level of detail.

For our analysis of monosyllables, we extracted all monosyllabic nouns, verbs and adjectives that were not potential outputs of OSL. Then, we categorized them with regard to their metrical weight and prosodic structure into words of the *land-type* (heavy, ending in a consonant cluster), the *mood-type* (heavy, long vowel) and the *god-type* (light, short vowel). Finally, we determined their vowel height (high vs. non-high) and counted their type and token frequencies.

For disyllables, we proceeded similarly, except that we worked with a sample rather than the complete set of attested forms. Once again, we extracted nouns, verbs and adjectives that were not potential OSL outputs and categorized them according to the structure of their first syllables into words of the *finger-type* (heavy, closed first syllable), the *bailiff-type* (heavy, long first syllable) and the *mother-type* (light, short first syllable). In addition, we determined their morphological structure (simple or complex). Then, we counted type and token frequencies of each combination of weight/prosodic structure and morphological structure.

3.1.1 Monosyllables

For our data set of monosyllables, we extracted from the LAEME corpus all monosyllabic word forms on the basis of their spelling (6394 nouns, 8809 verbs, 2411 adjectives). This data needed additional processing: first, we excluded items with open syllables (such as *fe* 'fee, livestock', *dai* 'day', or *fa* 'foe') because these were not comparable to OSL outputs, which were all closed. Second, we excluded items whose coda was an inflectional suffix (such as *see+s* 'sea.PL', *sai+d* 'say.PT', *seo+ð* 'see.3SGPRES', or *ga+n* 'go.INF'). Third, we excluded grammaticalized high frequency items because – due to their grammaticalization – they were not prototypical representatives of their word classes and were hypothesized not to serve as prototypical mental templates for newly emerging word shapes. The items excluded were the noun *man* (which also functioned as an indefinite pronoun), forms of *be, have*, or *do*, the modal verbs *may*, *will*, *shall*, *can*, the numeral adjectives *all*, *each*, *such*, *some*, and *which*. Finally, we excluded items that represented monosyllabic spellings of early OSL outputs, such as *nom* 'name', *sac* 'sake', or *meet* 'meat'. Potentially monosyllabic OSL outputs that were still spelt with final <e> or other vowel graphs were of course also excluded. After this processing, our final dataset for monosyllables included 2612 nouns, 1606 verbs and 735 adjectives.

3.1.2 Disyllables

For our dataset of disyllables, we first extracted all disyllabic word forms (33693 nouns, 39642 verbs, 12630 adjectives), and selected pseudo-random samples of 2000 nouns, verbs and adjectives each. We made sure that the mix of items with high and low token frequency in our samples reflected the token frequency mix of the complete dataset, except that we excluded *hapax legomena*. From our samples, we then excluded remaining items with transcription errors and items with ambiguous morphological structure, syllable weight or vowel length. Also, we excluded items whose final syllable was <-e> , since we could not determine if in our target period (1150-1325) final *-e* was still pronounced as /e/, reduced to schwa, or already lost completely. Also, exactly these items were the inputs of OSL, which may have been shaped by the majority patterns that we were interested in. Therefore, it would be unjustified to include these items in our dataset. The remaining dataset included 925 nouns, 1134 verbs and 700 adjectives that entered our analysis.

3.2 Data preparation and qualitative analysis

3.2.1 Preliminary remarks

The LAEME corpus provided us with lists of Early Middle English word forms attested in written texts. For each attested spelling variant, LAEME provided a lemma, a morpho-syntactic tag, and the token frequencies of this variant. The phonological information we needed to derive from the written forms, was (a) syllable boundaries, (b) syllable weight, (c) the phonological length of vowels, and (d) the height of vowels if they were monophthongs. The morphological information that we needed to derive for disyllabic items was whether they were simple or complex.

Inevitably, our categorizations involved a substantial degree of philological interpretation and were not always straightforward. This is because spelling does not represent pronunciation faithfully, and Middle English spelling was particularly variable. Also, the large number of examples we had to characterize made it impossible to consider all aspects that a careful philological interpretation would normally require. So, not all our categorizations may stand up to close philological scrutiny. However, in cases where we found it difficult to decide between alternative interpretations, we tried to make sure to settle for the one that was less favourable to our predictions, in order to counteract the effects of a possible confirmation bias.

In the following, we describe and illustrate the basic principles we applied in our analysis. A detailed discussion of the decisions that we made during our categorizations, including further examples, can be found in the supplementary material.

3.2.2 Syllable boundaries and syllable weight

To determine syllable weight, we identified syllable boundaries (in the case of disyllables) and syllable codas. For that, we interpreted consonant graphs as representing phonological consonants more or less faithfully. On that basis, we took syllabification to be onset maximal. Thus, we would syllabify a form such as *knictes* 'knight.PL' as *knic.tes*, and a form like *bagges* 'bag.PL' as *ba.gges*.

On the basis of these syllabifications, we determined syllable weight. In the case of monosyllables, we counted all syllables as heavy that had more than a single coda consonant, irrespectively of the length of their vowel (e.g. *mauht* 'might' or *milc*, 'milk'). In the case of disyllables, a single coda consonant in the first syllable counted as sufficient for making this syllable heavy (e.g. *knic.tes* 'knights', *al.mes* 'alms', or *an.gel* 'angel'). For more detailed information, see the supplementary material.

3.2.3 Vowel quantity and vowel height

We determined vowel length and vowel height, i.e. vowel quality, by considering vowel length and quality in Old English or Modern English reflexes, and by consulting dictionaries such as the Oxford English Dictionary (https://oed.com/), the Middle English Dictionary (https://quod.lib.umich.edu/m/middle-english-dictionary/dictionary), or the Dictionary of Old English (https://tapor.library.utoronto.ca/doe/). For more detailed information, see the supplementary material.

3.2.4 Morphotactic analysis

For the morphotactic analysis we could rely on the grammatical tags provided in LAEME. For example, the form *comeð* 'come' is tagged as a second person plural imperative. Since the imperative has an evident phonological exponent, namely *-eð*, we confidently classified *comeð* as morphologically complex, and proceeded in the same way with all other cases.

3.3 Quantitative data analysis

To analyse the proportions of different phonotactic shapes and morphological patterns, we counted both type and token frequencies. Our basis for establishing what should count as a single type were unique combinations of sound shape, lemma and grammatical tag. For example, the seven spellings of *land* in Tab. 5 counted as two different types because – even though they shared the same sound shape and represented the same lemma – three spellings represented the nominative form, and four the oblique form (n>pr, i.e. noun forms preceded by prepositions). For token frequencies, we used those reported in LAEME.

Table 5. Classification of types with regard to spelling, sound shape, lemma and grammatical tag.

Type	Spelling	Sound shape	Lemma	Tag
1	*lond*	CVCC	*land*	n
	long	CVCC	*land*	n
	lont	CVCC	*land*	n
2	*lond*	CVCC	*land*	n<pr
	land	CVCC	*land*	n<pr
	loand	CVCC	*land*	n<pr
	lonð	CVCC	*land*	n<pr

To compare the proportions of phonotactic shapes and to establish which of them represented the majority, we calculated 95% confidence intervals. Confidence intervals that do not overlap with one another indicate significant differences between groups. Additionally, confidence intervals that do not include the 50% mark indicate that a pattern is either in the majority (above 50%), or in the minority (below 50%; Cumming 2014; Cumming and Finch 2005; Cumming 2012).

For disyllables, we additionally operationalized the relationship between morphological structure and sound shapes by calculating chi-squared tests and phi correlation coefficients, which measure the correlation between two binary variables (Everitt and Skrondal 2010; Warrens 2008; Yule 1912). A phi coefficient of 1 indicates a perfect correlation between morphological structure and sound shapes. This would be the case, for example, if all morphologically complex word forms had long vowels in their first syllables and all morphologically simple word forms short vowels, or *vice versa*. A phi coefficient of 0 indicates that there is no clear relationship between morphological structures and sound shapes and that listeners will be unable to infer morphological structure from sound shapes and *vice versa*. Commonly, phi coefficients around 0.3 indicate medium correlations and phi coefficients around 0.5 strong correlations (Cohen 1992). – All calculations were done in *R* (version 3.6.0; *R* Development Core Team 2018).

4 Results: monosyllables

4.1 Syllable weight

Our analyses revealed that the proportions of heavy Early Middle English monosyllabic nouns, verbs and adjectives clearly lay above 50%. This was true for word form types (nouns: 81.16%, verbs: 83.75%, adjectives: 81.23%; CIs do not include 50%; Fig 3a) and for word tokens (nouns: 86.47%, verbs: 77.17%, adjectives: 86.72%; CIs do not include 50%; Fig. 3b). This means that the clear majority of Early Middle English monosyllables was heavy. Therefore, heavy monosyllables were much more probable as representatives of monosyllabic words than light monosyllables, which matches prediction 1 (Fig. 1), our weakest prediction about monosyllables.

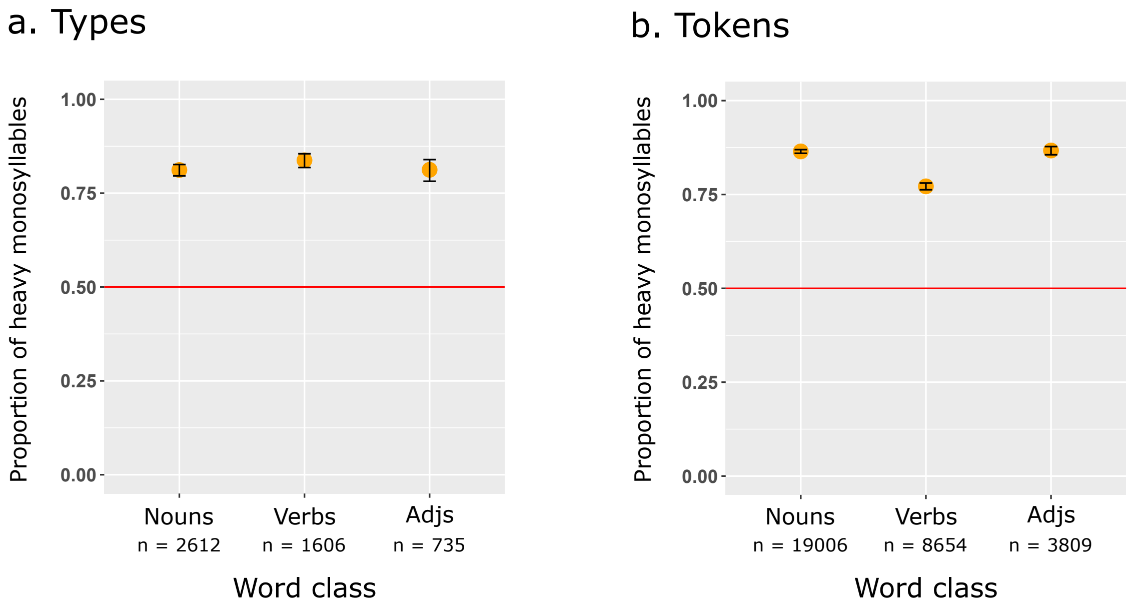


Figure 3. Proportions of heavy monosyllable (a) types and (b) tokens in Early Middle English nouns, verbs and adjectives. Error bars represent 95 % confidence intervals. The red line indicates 50%, i.e. proportions at which sound patterns are maximally ambiguous.

4.2 Vowel length

Our analyses revealed that the clear majority of monosyllabic nouns, verbs and adjectives that ended in single consonants had long vowels or diphthongs, and only a minority had short vowels. Again, these relations held for word form types (nouns: 67.93%, verbs: 68.97%, adjectives: 70.07%; CIs do not include 50%; Fig 4a) and word tokens (nouns: 71.14%, verbs: 60.40%, adjectives: 78.29%; CIs do not include 50%; Fig. 4b). Thus, words from the *mood*-type were more typical representatives of monosyllabic words than words from the *god*-type, which matches prediction 2 (Fig. 1), our stronger predictions about monosyllables.

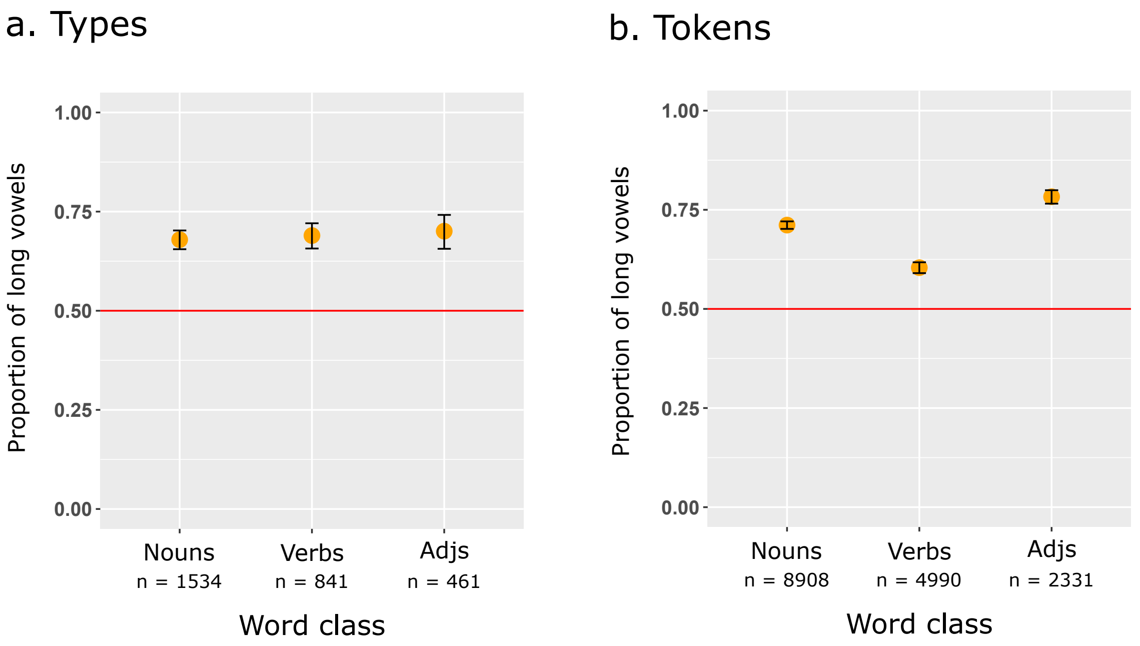


Figure 4. Proportions of monosyllable (a) types and (b) tokens with long vowels or diphthongs in Early Middle English nouns, verbs and adjectives. Error bars represent 95 % confidence intervals. The red line indicates 50%, i.e. proportions at which sound patterns are maximally ambiguous.

4.3 Vowel length in high vs. non-high monosyllables

A comparison between Early Middle English monosyllables with high and non-high vowels revealed also clear differences between these groups, which are roughly in line with prediction 3 (Fig. 1), our strongest prediction for monosyllables. The prediction is met unambiguously insofar as the clear majority of non-high vowels was long (types: nouns: 69.60%, verbs: 71.65%, adjectives: 74.73%; CIs do not include 50%; Fig. 5a; tokens: nouns: 71.87%, verbs: 64.22%, adjectives: 79.06%; CIs do not include 50%; Fig. 5b). However, our prediction that the majority of high vowels would be short, was borne out only partly. On the type level, the proportion of short high vowels lay around 50% for verbs and adjectives, and in the case of nouns, a narrow majority of high vowels was in fact long (nouns: 55.87%, verbs: 51.04%, adjectives: 55.13%; CIs of verbs and adjectives include 50%; Fig. 5a). On the level of tokens, the majority of vowels was short only for verbs, but not for nouns and adjectives (nouns: 61.79%, verbs: 31.86%, adjectives: 63.72%; Fig. 5b). Nevertheless, the proportion of long vowels was always significantly greater among non-high vowels than among high ones (see non-overlapping confidence intervals in Fig. 5 for noun, verb and adjective types and tokens), which is why, overall, our strongest prediction about monosyllables can still be considered to be borne out.

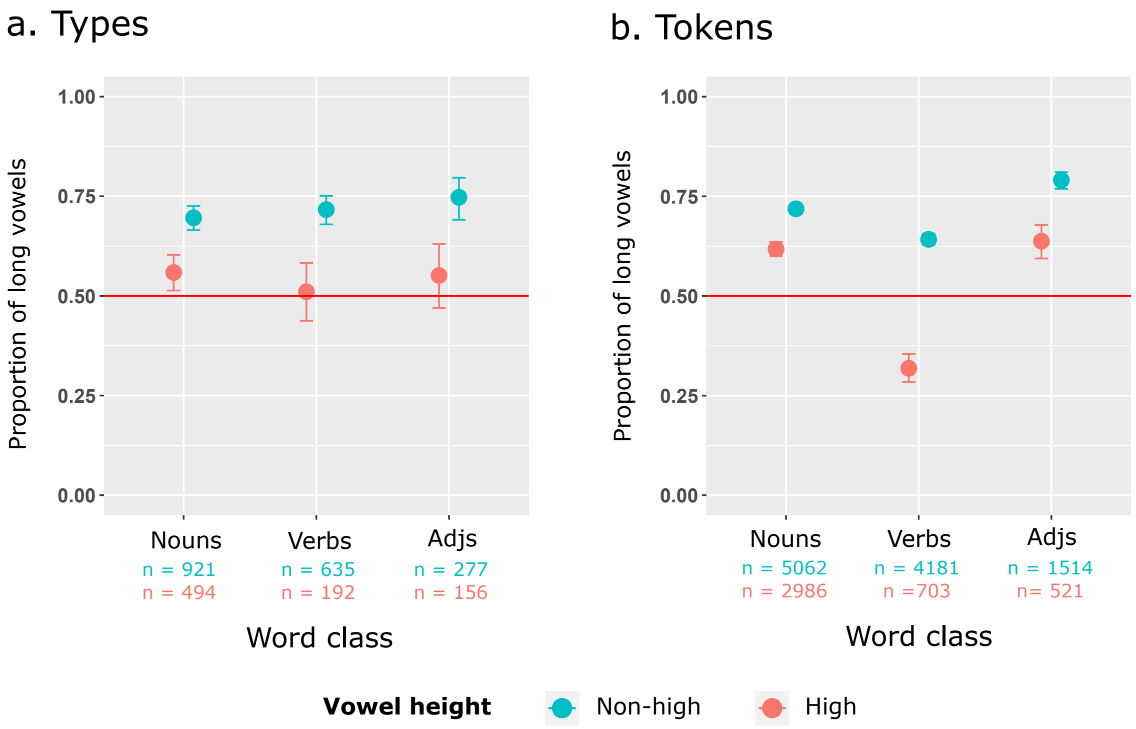


Figure 5. Proportions of monosyllable (a) types and (b) tokens with long vowels in Early Middle English nouns, verbs and adjectives, distinguished by vowel height. Error bars represent 95 % confidence intervals. The red line indicates 50%, i.e. proportions at which sound patterns are maximally ambiguous.

5 Results: disyllables

5.1 Proportions of simple and complex items among disyllables with different shapes of their first syllable

5.1.1 Syllable weight

Among disyllabic nouns and adjectives, the majority of word form types with light first syllables were morphologically simple. This was true both for types (nouns: 59.19%, adjectives: 64.35% simple; CIs do not include 50%; Fig. 6a) and for tokens (nouns: 62.98%, adjectives: 85.18% simple; CIs do not include 50%; Fig. 6b). In contrast, the majority of word form types with heavy initial syllables were complex. This also held on the level of types (nouns: 25.16%, adjectives: 20.86% simple; CIs do not include 50%; Fig. 6a) and tokens (nouns: 24.79%, adjectives: 45.62% simple; CIs do not include 50%; Fig. 6b). The medium to strong correlations between morphological structure and initial syllable weight in nouns and adjectives (see results of chi-squared tests and phi-correlation coefficients in Tab. 6 and 7) further support the significance of these relationships. – Thus, our predictions 1 and 2 for disyllables (Fig. 2) were borne out well among nouns and adjectives: disyllables with heavy open first syllables were more often complex than simple, and disyllables with light open first syllables were more often simple than complex.

Since our dataset did not include a sufficient number of morphologically simple verbs (2 types and 4 tokens), no conclusions about verbs can be drawn.

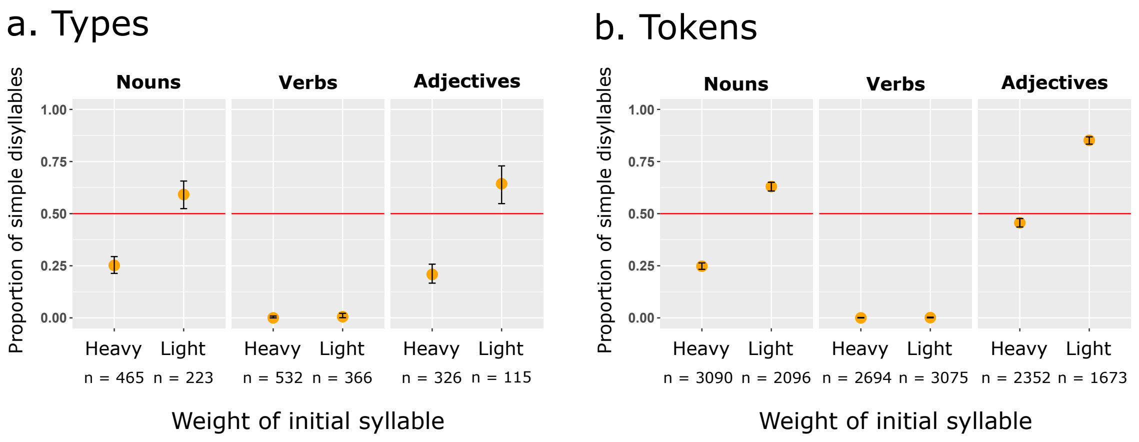


Figure 6. Proportions of morphologically simple word forms in Early Middle English disyllable noun, verb and adjective (a) types and (b) tokens with heavy and light initial syllables. Error bars represent 95 % confidence intervals. The red line indicates 50%, i.e. proportions at which sound shape patterns are least indicative for morphological structures.

Table 6. Number of word form types with heavy and light initial syllables in morphologically simple and complex nouns, verbs and adjectives.

Word class	Weight	Simple	Complex	Correlation
Nouns	Heavy	117	348	$\chi^2 = 74.12$, $p < 0.001$, φ = 0.33
	Light	132	91	
Verbs	Heavy	0	532	$\chi^2 = 0.97$, $p = 0.324$, φ = 0.03*
	Light	2	364	
Adjectives	Heavy	68	258	$\chi^2 = 71.67$, $p < 0.001$, φ = 0.40
	Light	74	41	

* Correlation of limited interpretability because of the low samples size of simple items in our dataset

Table 7. Number of word tokens with heavy and light initial syllables in morphologically simple and complex nouns, verbs and adjectives.

Word class	Weight	Simple	Complex	Correlation
Nouns	Heavy	766	2324	$\chi^2 = 755.85$, $p < 0.001$, φ = 0.38
	Light	1320	776	
Verbs	Heavy	0	2694	$\chi^2 = 1.88$, $p = 0.170$, φ = 0.02*
	Light	4	3071	
Adjectives	Heavy	1073	1279	$\chi^2 = 647.98$, $p < 0.001$, φ = 0.40
	Light	1425	248	

* Correlation of limited interpretability because of the low samples size of simple items in our dataset

5.1.2 Vowel length

According to our prediction, among disyllabic nouns and adjectives with short vowels in open first syllables, the majority were simple. This held among both types (nouns: 59.19%, adjectives: 64.34% simple; CIs do not include 50%; Fig. 7a) and tokens (nouns: 62.98%, adjectives: 85.18% simple; CIs do not include 50%; Fig. 7b). These numbers are identical to those for disyllables with light first syllables (Fig. 6) because light syllables are per definition identical to open short ones.

Among disyllables with long vowels in open first syllables, there was a difference between the type level and the token level. On the type level, both nouns and adjectives were distributed as we predicted. On the type level, the majority of both adjectives and nouns with long vowels in their open first syllable were complex (nouns: 18.65%, adjectives: 30.06% simple; CIs do not include 50%; Fig. 7a). On the token level, however, adjectives differed from nouns. Among nouns, the majority of items with long vowels in open first syllables were complex (14.95% simple; CIs do not include 50%; Fig 7b). Among adjectives, however, the majority were simple (63.32% simple; CIs do not include 50%; Fig 7b), although that majority was not as great as among adjective tokens with short vowels in open first syllables. – In spite of the odd behaviour of adjective tokens, however, the relationships between short and long vowels in open first syllables and complex vs. simple word forms displayed medium to strong correlations between morphological structure and initial vowel length in nouns and adjectives (see results of chi-squared tests and phi-correlation coefficients in Tab. 8 and 9). Thus, our predictions 3 and 4 for disyllables (Fig. 2) were on the whole borne out well: disyllables with long vowels in open first syllables were more often complex than simple, and disyllables with short vowels in open first syllables were more often simple than complex. Once again, it has to be pointed out, that the low number of morphologically simple verb forms, did not allow us to draw any conclusions.

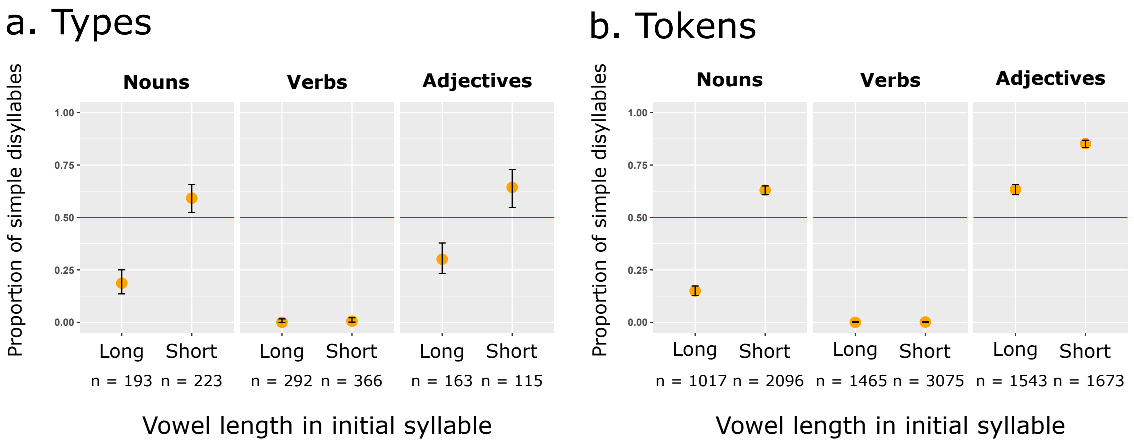


Figure 7. Proportions of morphologically simple word forms in Early Middle English disyllable noun, verb and adjective (a) types and (b) tokens with long and short initial syllables. Error bars represent 95 % confidence intervals. The red line indicates 50%, i.e. proportions at which sound shape patterns are least indicative for morphological structures.

Table 8. Number of word form types with long and short vowels in their initial syllables in morphologically simple and complex nouns, verbs and adjectives. Note that the data for items with short initial vowels is identical to the data for light initial syllables in Tab. 6.

Word class	**Vowel length**	**Simple**	**Complex**	**Correlation**
Nouns	Long	36	157	

	Short	132	91	$\chi^2 = 68.95$, $p < 0.001$, $\varphi = 0.41$
Verbs	Long	0	292	$\chi^2 = 0.30$, $p = 0.581$, $\varphi = 0.02$*
	Short	2	364	
Adjectives	Long	49	114	$\chi^2 = 30.76$, $p < 0.001$, $\varphi = 0.33$
	Short	74	41	

* Correlation of limited interpretability because of the low samples size of simple items in our dataset

Table 9. Number of word tokens with long and short vowels in their initial syllables in morphologically simple and complex nouns, verbs and adjectives. Note that the data for items with short initial vowels is identical to the data for light initial syllables in Tab. 7.

Word class	**Vowel length**	**Simple**	**Complex**	**Correlation**
Nouns	Long	152	865	$\chi^2 = 631.83$, $p < 0.001$, $\varphi = 0.45$
	Short	1320	776	
Verbs	Long	0	1465	$\chi^2 = 0.72$, $p = 0.398$, $\varphi = 0.01$*
	Short	4	3071	
Adjectives	Long	977	566	$\chi^2 = 201.71$, $p < 0.001$, $\varphi = 0.25$
	Short	1425	248	

* Correlation of limited interpretability because of the low samples size of simple items in our dataset

5.2 Proportions of different first-syllable shapes among simple and complex disyllables

5.2.1 Syllable weight

Among morphologically complex disyllabic adjectives and nouns, the clear majority of word forms had heavy first syllables. This was true both for types (nouns: 79.27%, adjectives: 68.29% heavy; CIs do not include 50%; Fig. 8a) and for tokens (nouns: 74.97%, adjectives: 83.76% heavy; CIs do not include 50%; Fig. 8b). Among morphologically simple disyllabic adjectives and nouns, the proportions of heavy first syllables were lower. For types, the proportions of heavy first syllables were around 50% (nouns: 46.99%, adjectives: 47.89% heavy; CIs include 50%; Fig. 8a) and for tokens, they were below 50% (nouns: 36.72%, adjectives: 42.95% heavy; CIs do not include 50%; Fig. 8b). For nouns and adjectives, this relationship is also reflected in significant medium to strong correlations between the morphological structure of words and the weight of their first syllables (see results of chi-squared tests and phi-correlation coefficients in Tab. 6 and 7).

Among disyllabic verbs, the majority of complex word form types (59.38% heavy; Fig. 8a) had heavy first syllables. However, among complex verb tokens the proportion of heavy first syllables was slightly below 50% (46.73% heavy; Fig. 8b). About simple disyllabic verbs, nothing can be said because there were hardly any of them in our sample (2 types, and 4 tokens, which is not surprising since verbal inflection was still intact in Early Middle English). – Overall, our data match our prediction 6 for disyllables (Fig. 2) well: first syllables were more often heavy in complex word forms. Prediction 5 (Fig. 2) is also met, but not as clearly: first syllables are indeed more often light in simple word form tokens, but for simple word form types, the proportion of light syllables lies just around 50%.

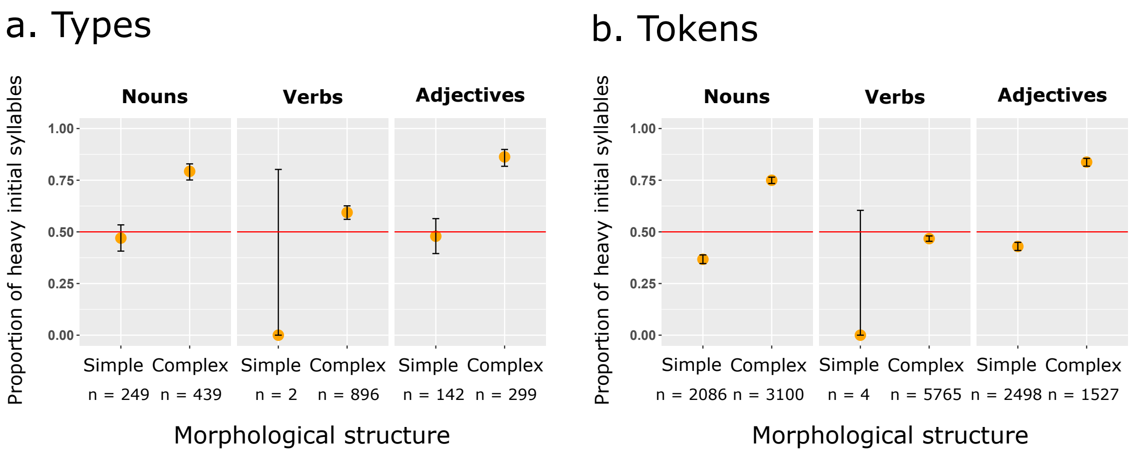


Figure 8. Proportions of disyllabic word forms with heavy initial syllables in morphologically simple and complex Early Middle English noun, verb and adjective (a) types and (b) tokens. Error bars represent 95 % confidence intervals. The red line indicates 50%, i.e. proportions at which morphological structures are least indicative for sound shapes.

5.2.2 Vowel length

For disyllables with open first syllables, our results were very similar to those for syllable weight. Among complex disyllabic nouns and adjectives with open first syllables, the clear majority had long vowels in their first syllable. This was true both for types (nouns: 63.31%, adjectives: 73.55% long; CIs do not include 50%; Fig. 9a), and for tokens (nouns: 52.71%, adjectives: 69.53% long; CIs do not include 50%; Fig. 9b). In contrast, the majority of simple word form types had short vowels in first syllables. Once again, this held for types (nouns: 21.43%, adjectives: 39.84% long; CIs do not include 50%; Fig. 9a) and tokens (nouns: 10.32%, adjectives: 40.67% long; CIs do not include 50%; Fig. 9b).

These relations also manifest in significant medium to strong correlations between morphological structure and vowel length in word-initial syllables (see results of chi-squared tests and phi-correlation coefficients in Tab. 8 and 9). Thus, among nouns and adjectives, our predictions 7 and 8 (Fig. 2) were borne out well: Among complex disyllables, the majority of first syllables were long rather than short, while among simple disyllables, the opposite was true.

Once again, the picture is less clear for verbs. In contrast to nouns and adjectives, the majority of complex disyllabic verbs had short vowels in their first syllables, although for word form types, the proportion of short vowels in initial syllables lies only marginally above 50% (verb types: 44.51%; verb tokens: 32.30% long; Fig. 9). Again, we cannot say anything about simple verbs.

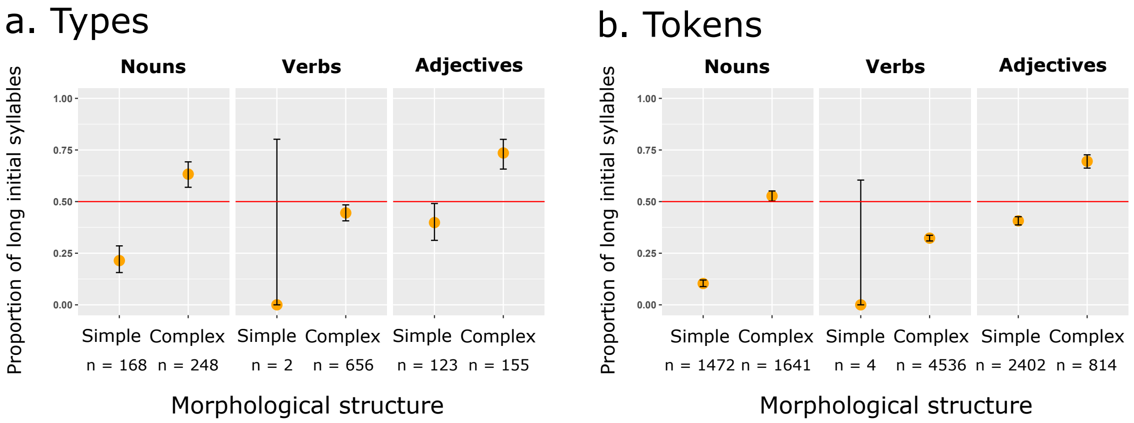


Figure 9. Proportions of disyllabic word forms with long initial syllables in morphologically simple and complex Early Middle English noun, verb and adjective (a) types and (b) tokens. Error bars represent 95 % confidence intervals. The red line indicates 50%, i.e. proportions at which morphological structures are least indicative for sound shapes.

6 Discussion

Evidently, our results see practically all our predictions borne out: at the time when OSL set in, long and short vowels and heavy and light syllables were distributed among monosyllabic and disyllabic word forms so that the way in which OSL was implemented stabilized or even increased the probability of word form shapes that were already in the majority. Thus, the regular lengthening of non-high vowels in the *make-type* increased the probability of monosyllabic word forms to be heavy rather than light, and to have long rather than short vowels if they did not end in consonant clusters. Likewise, the failure of high vowels to lengthen in such cases matches the fact that high vowels in Early Middle English monosyllables were typically not more often long than short.

Among disyllables, the failure of most vowels to undergo OSL corresponds to the fact that the majority of simple disyllables had short vowels in light first syllables at the time when OSL set in. In contrast, heavy first syllables (no matter of closed or open) were more frequent among complex disyllabic word forms. That relationship also held the other way round: if a disyllabic word had a light first syllable (i.e. a short vowel), then it would have been simple in the majority of cases, and if its first syllable was heavy it would have been complex. Thus, the distribution of long and short vowels among disyllables was a good indicator of their morphological structure. The implementation of OSL not only helped to maintain this relationship by not lengthening vowels in words of the *habit-type*, but it even increased that indicativeness further, albeit indirectly, by lengthening vowels in words that became monosyllabic. This is because any inflected forms of such words (e.g. *makeð* ‘make 3.SGPRES’, or ‘*names* ‘name PL’) would increase the already high probability of disyllables with heavy first syllables to be complex. Our findings go hand in hand with a previous study showing that the voicing of the final /s/ in English noun plurals had the effect of making these complex forms easy to distinguish from morphologically simple items, among which final /z/ is very rare (Baumann, Prömer, and Ritt 2019).

We take our results to be quite spectacular. It seems as if the distributions of heavy and light syllables (and long and short vowels) among lexical monosyllables and disyllables would predict the eventual implementation of OSL almost exactly, almost as if no other phonological conditions were involved.[4] Of course, the correlations we have found are just that: namely correlations. At the same time, they are so strong and so specific that they make it worthwhile to discuss causalities they might reflect.

Like most sound changes, OSL is likely to have started on the phonetic level, by lengthening the duration of vowels that were phonologically short. Thus, their duration would have become ambiguous as an indicator of the intended short vowels, and these vowels may have been reinterpreted as reflecting phonologically long ones. So, phonologically long and short variants will have competed as phonological representations of OSL inputs. If words, and word forms with probable phonotactic shapes are easier to identify and to learn than words with less probable shapes, this may have selected for those variants that did have the more probable and morpho-syntactically more indicative shapes. Among words of the *make-type* these were the variants with long vowels, and among words of the *beaver/habit-types* they were the ones with short vowels. As far as we see it, such an account would be logically consistent, and there is much independent evidence for all the processes and preferences it needs to appeal to – both from socio-historical phonology and from psycholinguistics. In particular, such an account would clearly be compatible with, and support the general hypothesis that a preference for word forms to assume probable shapes represents a possibly universal cognitive bias that may interact with other factors to constrain the evolution of sound patterns (Bybee 2007; Ambridge et al. 2015; Diessel 2007; Ellis 2002; Divjak 2019a, 2019b).

Of course, the way in which preferences based on lexical statistics interact with other factors may be complex. Consider for example the case of high vowels in monosyllables. While their failure to implement OSL seems to be predictable by the fact that long items like *wif* 'wife' /wi:f/ or *house* /hu:s/ were not significantly more probable at the time when OSL set in than short items like *wit* or *full*, it may equally well have been caused by the inherently shorter duration of high vowels in comparison to non-high ones (House 1961; Delattre 1962; Lehiste 1970; Lisker 1974). Indeed, the inherent shortness may underlie both the relative rarity of *wif*-type words and the failure of high vowels to undergo OSL at the same time. However, even if that should be the case, the two factors may have mutually supported one another.

More generally speaking, the potential importance of lexical probabilities, which our findings suggest, does not invalidate the importance of other phonological factors. These may be the open syllable condition itself, the quality of the postvocalic consonants, or the structure of the second syllable if it was retained. Since our focus has been on the potential impact of lexical probability, we have not discussed the details of these phonological conditions on OSL (see e.g. Ritt 1994; Minkova 1982; Bermudez-Otero 1998; Mailhammer, Kruger, and Makiyama 2015; Minkova and Lefkowitz 2020; Minkova, forthc.; Lahiri and Dresher 1999 for in depth discussions). Therefore, our findings are clearly not intended to compete with extant accounts but rather to complement them.

A final aspect is that, overall, type level results were more strongly compatible with our predictions than token level results. This is plausible because it is compatible with similar insights on language acquisition and learning (e.g. Bybee 1995; Ellis 2002; Lieven 2010; Endress and Hauser 2011). The correlations we have demonstrated involve abstractions on a comparably high level, namely between syllabic structures that can be realized by a variety of different segment sequences, and morphotactic structures that can likewise be realized by a variety of different morpheme combinations. To learn that there is a statistical correlation

[4] Of course, we are aware that other phonological conditions played an additional role and can explain why lengthening occurred in the *beaver*-type but not in the *habit*-type (see Minkova & Lefkovitz 2020).

between abstract phonotactic patterns such as an initial heavy syllable and abstract complex morphotactic patterns such as stem+suffix is very likely to require exposure to many different types of these patterns. A few types may not be enough, even if they are highly frequent in terms of tokens. Thus, the fact that type-level results show clearer correlations than token-level results is not surprising.

7 Limitations, conclusion and outlook

Although practically all our predictions have been borne out, it needs to be stressed that they merely support the plausibility of the general hypothesis that lexical probabilities may constrain the implementation of sound changes. They clearly do not prove it. Among other things this is because our argumentation has been abductive. We have started from the observation that OSL was regularly implemented among disyllables that had non-high vowels and became monosyllabic, and that it was implemented only rarely among words with high vowels and among words that retained their second syllable. We then defined the conditions under which a preference for words with probable and morpho-syntactically indicative sound shapes would predict the attested implementation pattern, and finally we enquired if these conditions held. That we did indeed find the necessary conditions to hold, therefore merely suggests that our hypothesis is plausible but does not prove that the causalities it implies were really involved in producing the attested implementation pattern.

However, despite such limitations we take our findings to be interesting enough to warrant further research. In particular, and even though our results concern only a single and quite specific case of a sound change, they suggest that lexical probabilities may play a greater role in the actuation and the implementation of phonological change than currently known. Given the increasing availability of digitized corpora and dictionaries of historical language stages, investigations of such a role may become more practicable than they have been and could also be extended to languages beyond English and phenomena beyond OSL. Such research could further support that sound changes are more likely to be actuated and implemented if they stabilize or increase the probability of already probable sound patterns, which would considerably advance our understanding of phonological evolution.

Acknowledgements

We would like to thank xxx, xxx, xxx and xxx for helpful feedback on the project.

Data availability statement

The datasets generated and analysed during the current study (doi: 10.17605/OSF.IO/CKMSH) are available in the Open Science Framework repository and can be accessed at https://osf.io/ckmsh/?view_only=d26e756e78064b659ef236be4dfadd7f.

References

Ambridge, Ben, Evan Kidd, Caroline F. Rowland, and Anna L. Theakston. 2015. “The Ubiquity of Frequency Effects in First Language Acquisition.” *Journal of Child Language* 42: 239–73.

Baumann, Andreas, and Kamil Kaźmierski. 2018. “Assessing the Effect of Ambiguity in Compositionality Signaling on the Processing of Diphones.” *Language Sciences* 67 (May): 14–32. https://doi.org/10.1016/j.langsci.2018.03.006.

Baumann, Andreas, Christina Prömer, and Nikolaus Ritt. 2019. “Word Form Shapes Are

Selected to Be Morphotactically Indicative." *Folia Linguistica* 40 (1): 129–51. https://doi.org/10.1515/flih-2019-0007.

Berg, Thomas. 2014. "On the Relationship between Type and Token Frequency." *Journal of Quantitative Linguistics* 21 (3): 199–222. https://doi.org/10.1080/09296174.2014.911505.

Bermudez-Otero, Ricardo. 1998. "Prosodic Optimization: The Middle English Length Adjustment." *English Language and Linguistics* 2 (02): 169–97. https://doi.org/10.1017/S1360674300000848.

Blevins, Juliette. 2006. "A Theoretical Synopsis of Evolutionary Phonology." *Theoretical Linguistics* 32 (2): 117–166. https://doi.org/10.1515/TL.2006.009.

———. 2009. "Structure-Preserving Sound Change: A Look at Unstressed Vowel Syncope in Austronesian." In *Austronesian Historical Linguistics and Culture History: A Festschrift for Bob Blust*, edited by Alexander Adelaar and Andrew Pawley, 33–49. Canberra: Pacific Linguistics.

Bybee, Joan. 1995. "Regular Morphology and the Lexicon." *Language and Cognitive Processes* 10 (5): 425–55. https://doi.org/10.1080/01690969508407111.

———. 2007. *Frequency of Use and the Organization of Language*. Oxford: Oxford University Press.

Calderone, Basilio, Chiara. Celata, Katharina Korecky-Kröll, and Wolfgang Ulrich Dressler. 2014. "A Computational Approach to Morphonotactics: Evidence from German." *Language Sciences* 46: 59–70.

Cohen, Jacob. 1992. "A Power Primer." *Psychological Bulletin* 112 (1): 155–59.

Cole, Jennifer, Yoonsook Mo, and Mark Hasegawa-Johnson. 2010. "Signal-Based and Expectation-Based Factors in the Perception of Prosodic Prominence." *Laboratory Phonology* 1 (2): 425–52. https://doi.org/10.1515/labphon.2010.022.

Cumming, Geoff. 2012. *Understanding The New Statistics*. New York: Routledge.

———. 2014. "The New Statistics: Why and How." *Psychological Science* 25 (1): 7–29. https://doi.org/10.1177/0956797613504966.

Cumming, Geoff, and Sue Finch. 2005. "Inference by Eye Confidence Intervals and How to Read Pictures of Data." *American Psychologist* 60 (2): 170–80. https://doi.org/10.1037/0003-066X.60.2.170.

Delattre, Pierre. 1962. "Some Factors of Vowel Duration and Their Cross-Linguistic Validity." *The Journal of the Acoustical Society of America* 34: 1141–43.

Diessel, Holger. 2007. "Frequency Effects in Language Acquisition, Language Use, and Diachronic Change." *New Ideas in Psychology* 25: 108–27. https://doi.org/10.1016/j.newideapsych.2007.02.002.

Divjak, Dagmar. 2019a. "Counting Occurrences: How Frequency Made Its Way into the Study of Language." *Frequency in Language*, 15–39. https://doi.org/10.1017/9781316084410.002.

———. 2019b. *Frequency in Language - Memory, Attention and Learning*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316084410.

Dressler, Wolfgang U, and Katarzyna Dziubalska-Kołaczyk. 2006. "Proposing Morphonotactics." *Wiener Linguistische Gazette* 73 (1–19): 108–9.

Ellis, Nick C. 2002. "Frequency Effects in Language Processing." *Studies in Second Language Acquisition* 24: 143–88.

Endress, Ansgar D., and Marc D. Hauser. 2011. "The Influence of Type and Token Frequency on the Acquisition of Affixation Patterns: Implications for Language Processing." *Journal of Experimental Psychology: Learning Memory and Cognition* 37 (1): 77–95. https://doi.org/10.1037/a0020210.

Everitt, B S, and A Skrondal. 2010. *The Cambridge Dictionary of Statistics*. 4th ed. Cambridge: Cambridge University Press.

https://ejournal.poltektegal.ac.id/index.php/siklus/article/view/298%0Ahttp://repositorio.unan.edu.ni/2986/1/5624.pdf%0Ahttp://dx.doi.org/10.1016/j.jana.2015.10.005%0Ahttp://www.biomedcentral.com/1471-2458/12/58%0Ahttp://ovidsp.ovid.com/ovidweb.cgi?T=JS&P.

Fehér, Olga, Elizabeth Wonnacott, and Kenny Smith. 2016. "Structural Priming in Artificial Languages and the Regularisation of Unpredictable Variation." *Journal of Memory and Language* 91: 158–80. https://doi.org/10.1016/j.jml.2016.06.002.

Gibson, Edward, Richard Futrell, Steven T. Piandadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. 2019. "How Efficiency Shapes Human Language." *Trends in Cognitive Sciences* 23 (5): 389–407. https://doi.org/10.1016/j.tics.2019.02.003.

Goldrick, Matthew, and Meredith Larson. 2008. "Phonotactic Probability Influences Speech Production." *Cognition* 107 (3): 1155–64. https://doi.org/10.1016/j.cognition.2007.11.009.

House, Arthur S. 1961. "On Vowel Duration in English." *The Journal of the Acoustical Society of America* 33 (9): 1174–78.

Kelley, Matthew C., and Benjamin V. Tucker. 2017. "The Effects of Phonotactic Probability on Auditory Recognition of Pseudo-Words." *The Journal of the Acoustical Society of America* 141 (5): 4038. https://doi.org/10.1121/1.4989319.

Kirby, Simon, Tom Griffiths, and Kenny Smith. 2014. "Iterated Learning and the Evolution of Language." *Current Opinion in Neurobiology* 28: 108–14. https://doi.org/10.1016/j.conb.2014.07.014.

Korecky-Kröll, Katharina, Wolfgang Ulrich Dressler, Eva Maria Freiberger, Eva Reinisch, Karlheinz Mörth, and Gary Libben. 2014. "Morphonotactic and Phonotactic Processing in German-Speaking Adults." *Language Sciences* 46: 48–58.

Lahiri, Aditi, and B Elan Dresher. 1999. "Open Syllable Lengthening in West Germanic." *Language* 75 (4): 678–719.

Laing, Margaret. 2013. "A Linguistic Atlas of Early Middle English, 1150-1325." Edinburgh: The University of Edinburgh.

Lange, Floris P. de, Micha Heilbron, and Peter Kok. 2018. "How Do Expectations Shape Perception?" *Trends in Cognitive Sciences* 22 (9): 764–79. https://doi.org/10.1016/j.tics.2018.06.002.

Lehiste, Ilse. 1970. *Suprasegmentals*. Cambridge: MIT Press.

Lieven, Elena. 2010. "Input and First Language Acquisition: Evaluating the Role of Frequency." *Lingua* 120 (11): 2546–56. https://doi.org/10.1016/j.lingua.2010.06.005.

Lisker, Leigh. 1974. "On 'Explaining' Vowel Duration Variation." *Glossa* 8 (2): 233–46.

Luick, Karl. 1964. *Historische Grammatik Der Englischen Sprache*. 1st ed. Leipzig: Tauchnitz.

Mailhammer, Robert, William W. Kruger, and Alexander Makiyama. 2015. "Type Frequency Influences Phonological Generalizations: Eliminating Stressed Open Syllables with Short Vowels in West Germanic." *Journal of Germanic Linguistics* 27 (3): 205–37. https://doi.org/10.1017/S1470542715000069.

McClelland, James L., and Jeffrey L. Elman. 1986. "The TRACE Model of Speech Perception." *Cognitive Psychology* 18 (1): 1–86. https://doi.org/10.1016/0010-0285(86)90015-0.

Minkova, Donka. n.d. "Preference Theory and the Uneven Progress of Degemination in Middle English."

———. 1982. "The Environment for Open Syllable Lengthening in Middle English." *Folia Linguistica Historica* 3 (1): 29–58.

———. 1991. *The History of Final Vowels in English: The Sound of Muting*. Vol. 4. Topics in English Linguistics. Berlin and New York: M. de Gruyter.

Minkova, Donka, and Michael Lefkowitz. 2020. "Middle English Open Syllable Lengthening

(MEOSL) or Middle English Compensatory Lengthening (MECL)?” *English Language and Linguistics*, no. December 2015: 1–26. https://doi.org/10.1017/S1360674319000522.

Pierrehumbert, Janet. 2016. “Phonological Representation: Beyond Abstract Versus Episodic.” *Annual Review of Linguistics* 2: 33–52.

Post, Brechtje, William D Marslen-Wilson, Billi Randall, and Lorraine K Tyler. 2008. “The Processing of English Regular Inflections: Phonological Cues to Morphological Structure.” *Cognition* 109 (1): 1–17. https://doi.org/10.1016/j.cognition.2008.06.011.

R Development Core Team. 2018. “R: A Language and Environment for Statistical Computing.” Vienna, Austria: R Foundation for Statistical Computing. http://www.r-project.org/.

Reali, Florencia, and Thomas L Griffiths. 2009. “The Evolution of Frequency Distributions : Relating Regularization to Inductive Biases through Iterated Learning.” *Cognition* 111 (3): 317–28. https://doi.org/10.1016/j.cognition.2009.02.012.

Richtsmeier, Peter T. 2011. “Word-Types, Not Word-Tokens, Facilitate Extraction of Phonotactic Sequences by Adults.” *Laboratory Phonology* 2 (1): 157–83.

Ritt, Nikolaus. 1994. *Quantity Adjustment: Vowel Lengthening and Shortening in Early Middle English*. Cambridge: Cambridge University Press.

Ritt, Nikolaus, and Kamil Kaźmierski. 2015. “How Rarities like Gold Came to Exist: On Co-Evolutionary Interactions between Morphology and Lexical Phonotactics.” *English Language and Linguistics*, no. 19: 1–29. https://doi.org/10.1017/S1360674315000040.

Smith, Kenny, Amy Perfors, Olga Fehér, Anna Samara, Kate Swoboda, and Elizabeth Wonnacott. 2017. “Language Learning, Language Use and the Evolution of Linguistic Variation.” *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 372 (1711): 1–20.

Storkel, H. L. 2001. “Learning New Words: Phonotactic Probability in Language Development.” *Journal of Speech, Language, and Hearing Research* 44 (6): 1321–37. https://doi.org/10.1044/1092-4388(2001/103).

Warrens, Matthijs J. 2008. “On Association Coefficients for 2x2 Tables and Properties That Do Not Depend on the Marginal Distributions.” *Psychometrika* 73: 777–89.

Wedel, Andrew. 2006. “Exemplar Models, Evolution and Language Change.” *The Linguistic Review* 23 (3): 247–74.

Yule, G Udny. 1912. “On the Methods of Measuring Association Between Two Attributes.” *Journal of the Royal Statistical Society* 75 (6): 579–652.

Word count excluding references: 8813 words

Phonotactically probable word shapes represent attractors in the cultural evolution of sound patterns

Supplementary material

1 Supplementary methods

1.1 Data preparation and qualitative analysis

For our analysis, we used data from the LAEME corpus, which includes Early Middle English word forms attested in written texts. We expanded this database by deriving syllable boundaries, syllable weight, vowel length and vowel height from the forms attested in the LAEME corpus. It needs to be pointed out that the purpose of our categorizations required us to apply abstractions, or idealisations, which paste over some of the fine distinctions that differences between individual spelling variants of a word from suggest. In that respect, our analysis does not do justice to the heterogeneity of different text languages that the LAME corpus represents. Instead, we have treated our data as representing something like an average Early Middle English, which represents an idealization that might not seem fully warranted from a variationist point of view but was required for the purposes of the present study.

The basic principles that we applied to prepare and classify our data are described in the methods section of the main paper. In this supplementary material, we illustrate our methodological decisions in more detail and provide additional examples for our classifications. Still, it would be impossible to discuss every single decision that we have made. Nevertheless, we are confident that the description will convey a sufficiently clear idea of our methods, and that the data base we prepared for quantitative analysis – while not free of errors – was not systematically biased in favour of our predictions.

1.1.1 Syllable boundaries and syllable weight

To determine the syllable weight of our target syllables, we identified syllable boundaries (in the case of disyllables) and the structure of syllable codas. For this task, it was relatively easy to rely on spelling evidence: we could interpret consonant graphs as representing phonological consonants more or less faithfully. Mostly, single consonant letters stood for single consonant phonemes, except in well-known digraphs such as <th>, which could stand for /ð/ or /θ/, <ch> or <cch>, which could stand for /tʃ/, etc. Also, we counted geminate graphs as single consonants. On that basis, we took syllabification to be onset maximal and did not allow ambisyllabicity. Thus, we would syllabify a form such as *knictes* 'knight.PL' as [*knic*]$_1$[*tes*]$_2$, rather than [*knic*[*t*]$_1$*es*]$_2$, and a form like *bagges* 'bag.PL' as [*ba*]$_1$[*gges*]$_2$, rather than [*bag*]$_1$[*ges*]$_2$.

As far as letters like <u>, <i>, <y>, <v>, and <w>, which could represent both consonants and vowels, were concerned, the decision was easy to make in the majority of cases. For example, in *deuel* 'devil', the <u> stood for a /v/, and we syllabified the from accordingly as [*de*][*uel*]. In ambiguous cases we chose a syllabication that was unfavourable to our predictions. For example, in the case of *thowth* 'thought', we could have counted the <w> as vocalic and analysed the form as CVVC, or as consonantal and have analysed the form as CVCC. Since our strongest prediction was that there should have been more CVVC items than CVC items in our data, we decided to analyse *thowth* as CVCC, in order to counteract the possible effects of a conformation bias.

The relative ease of interpreting consonant letters also facilitated the identification of heavy syllables. In the case of monosyllables, we classified all syllables as heavy that had more than a single coda consonant, irrespectively of the length of their vowel (e.g. *mauht* 'might', *mind*, or *milc*, 'milk'). In the case of disyllables, a single coda consonant counted as sufficient for making a syllable heavy (e.g. *al.mes* 'alms', *an.gel*, or *cast.les*). One of the measures we took to counteract confirmation bias was to assume that geminate consonants had simplified to single ones, both word-finally and also internally. Since vowels before originally geminate consonants were short, this increased the proportion of CVC items, which was unfavourable to our predictions.

1.1.2 Vowel quantity and vowel height

The determination of vowel length and vowel height, i.e. vowel quality, was a more challenging task since these properties are notoriously badly represented in spelling. Also, the spellings in LAEME reflect particularities of individual text languages, which we intended to abstract away from. What we tried to reconstruct were vowel quantities and vowel heights that were most probable at the time when OSL and schwa loss set in. We had to decide on them only in the case of monosyllables with single consonants in their codas, and disyllables with open first syllables, because all other items would be counted as heavy because of their codas, no matter whether their vowels were long or short.

The principles that we applied in our classification were the following: the easiest cases were items that had an Old English lemma in LAEME. This was the case for all words that had no Modern English descendants. In Old English lemmas, vowel quantity was marked, and we simply adopted the LAEME analysis. This happened for example in items such as *agt*, 'property', lemmatized as 'ae:ht', or *frið* 'peace, freedom', lemmatised as 'friY'.

In cases where Modern English reflexes suggested the same quantity as Old English ancestors (or reconstructible West Germanic ones), we assumed that the Middle English items would share that quantity. Straightforward examples would be *wif* 'wife', or *ban* 'bone', both with diphthongs in Modern English and long vowels in West Germanic (cf. MHG *Weib* and *Bein*).

Some cases were not quite as straightforward, but could be decided on the same principles. For example, although *blood* has a short vowel in Modern English, it can be reconstructed as deriving from an Old English long /o:/ that was raised to /u:/ in the Great Vowel Shift and must therefore have been long in Middle English as well. Thus, we classified forms like *blod* as having had a long vowel, in spite of their spelling.

Finally, in cases where our own knowledge of historical English phonology failed us, we relied on dictionaries such as the Oxford English Dictionary (OED: https://oed.com/), the Middle English Dictionary (MED: https://quod.lib.umich.edu/m/middle-english-dictionary/dictionary), or the Dictionary of Old English (DOE: https://tapor.library.utoronto.ca/doe/).

As pointed out, we tried to choose the unfavourable analysis when we were in doubt. For example, we analysed the vowel in *com* 'came.3SGPT' as short, because that was the quantity of inherited ablaut-grade of class IV verbs in the 1st and 2nd person singular of the past tense, even though it was likely that that form would have adopted a long vowel through analogical extension of the 2SG and PL forms.

In order to determine what vowels qualified as monophthongs, and what their height was, we relied on spelling, on our knowledge of historical English phonology, and on dictionaries. As far as spelling is concerned, the most problematic letter was <o>, which could represent both high /u(ː)/ (as in *son* ‘son’) and non-high /o(ː)/ or /ɔ:/ (as in *hom* ‘home’). This ambiguity also affected the diphthong spelling <ou> (/u:/ in forms such as *bour* ‘dwelling’, /ou/ in words like *douhter* ‘daughter’).

As in the case of vowel length, we tried to triangulate the correct interpretation from Modern English and Old English counterparts, and relied on dictionaries where we were in doubt. For distinguishing diphthongs from monophthongs, we proceeded accordingly. Broadly speaking, we classified vowel digraphs that ended in <i, y, u, v, w,> as rising diphthongs while we assumed Old English centring diphthongs to have monophthongised, so that we interpreted spellings such as <eo, ea, ie, etc.> as in *breost* ‘breast’, *breað* ‘breath’, or *dieð* ‘death’ as monophthongs.

Discussion

One of the core questions in the fields of comparative linguistics, cognitive linguistics, biolinguistics and animal communication research is how biology and culture interact during language acquisition, language change and language evolution. In this thesis, I explored what prosody as a means of structuring vocal signals can contribute to this issue. Specifically, the thesis provided insights into five questions at the intersection of biology and culture during language learning, change and evolution, namely a) how prosodic cues differ across non-human tetrapod vocalizations and various human languages, b) how successful different prosodic cues are for segmenting words from a continuous speech stream, c) how aesthetic different prosodic patterns are, d) how easily particular prosodic patterns are acquired in a second language, and e) how occurrence frequencies of prosodic patterns can influence how languages change diachronically. The following sections will outline what the findings of the thesis' five projects contribute to explaining the intersection of biological and cultural evolution during language transmission and will suggest promising avenues for future research.

1. Main findings of the thesis

The first project explored the fact that vocal production and perception mechanisms of non-human tetrapods are overall very similar to those of humans and that prosodic cues in animal and human vocalizations only differ in their details. This indicates that similar biological and cultural evolutionary pressures may have acted on the transmission of vocal signals across species and languages, and suggests that prosodic modifications that are similar in their acoustic realization across species and languages (such as final lengthening) may also be similar in the structural meaning they encode.

The second project found that different prosodic cues differed in how helpful they were for listeners when segmenting words from a continuous stretch of speech of an artificial language. Unsurprisingly, pauses between words were very salient for listeners and facilitated speech segmentation most. Interestingly, finally lengthened words were recognized almost as successfully as words separated by pauses, whereas finally shortened words were not segmented from continuous speech. Instead, shortened syllables were perceived as word-medial, even though this went against the typical stress pattern of the participants' native language (German). This suggests that listeners preferably used cross-linguistically consistent (and thus possibly physiologically rooted) boundary cues such as final lengthening to recognize words in continuous speech, whereas language-specific word-stress patterns were less important. Final pitch increase and decrease did not affect speech segmentation greatly, which points towards a more prominent role of durational modifications than pitch modifications for speech segmentation.

The third project found that different prosodic patterns differed in their aesthetic appeal. More precisely, finally lengthened words had a higher aesthetic appeal than finally shortened words. This is perfectly in line with the findings of the speech segmentation study from my second project, and points towards a possible link between language acquisition and aesthetic appeal. Also, as in the speech segmentation study, cross-linguistically consistent prosodic boundary cues had a special role: they were more aesthetically appealing than those prosodic patterns that were cues to language-specific word stress in German.

The fourth project indicated that non-native speakers and native speakers of English paused very similarly when reading an English text in different speech tempi. This suggests that salient structural cues such as pauses are acquired easily by second language learners, and that more cognitively demanding tasks such as speaking slowly or rapidly do not impede the successful production of these prosodic patterns. One explanation of this result may be that very general biologically driven production and processing biases led to highly similar pause patterns across languages during cultural evolution and that these patterns can be easily transferred between languages, even by second language speakers with noticeable foreign accents.

The fifth project found that one particular sound change in the history of English, Open Syllable Lengthening, adapted the prosodic shapes of newly established word forms to the shapes that were most frequent among already existing word forms. This suggests that highly frequent word shapes act as attractors during language change, biasing new words to become more similar to the majority of words in terms of their prosodic shapes. This may be because highly frequent forms are more characteristic and reliable, and have a cognitive processing advantage. In turn, this helps learners to segment the typical patterns more easily from continuous speech and to acquire and memorize them more successfully. The attracting effect of frequent patterns was not only found on the lexical but also on the morphological level, which shows that language learners are sensitive to fine-grained statistical regularities between prosodic patterns and structural meaning.

2. Contributions to non-human animal communication research and outlook to future research

Bioacoustics and non-human animal communication research provide an abundance of descriptive accounts of prosodic variation in non-human animal vocal signals (e.g. Morton, 2017). However, how exactly non-human animals perceive and interpret the voice modulatory cues that structure their vocal signals remains a major unresolved question in the field of non-human animal communication research.

My thesis identified starting points for investigating this question. One promising approach is to use prosodic patterns in human languages which listeners interpret as boundary signals cross-linguistically, such as final lengthening, as first candidates to test if non-human animals interpret these patterns similarly. The reason why particular prosodic patterns are interpreted similarly across dialects and languages may be that their production is determined by evolutionarily old physiological or energetic constraints, and as a result, these patterns signal certain structures unambiguously. For example, when stopping an utterance before a pause, it is physiologically easier to lengthen syllables before the pause instead of stopping abruptly (Edwards, Beckman, & Fletcher, 1991; Friberg & Sundberg, 1999), which makes final lengthening a very reliable and unambiguous boundary signal. Such unambiguous correspondences between signals and meanings are easily acquirable and transmissible, which may explain why prosodic patterns rooted in physiological biases are similar across dialects and languages. Because of their phylogenetic similarity to humans, it is plausible that similar mechanisms exist in non-human tetrapods (Toro & Crespo-Bojorque, 2021).

There is already some evidence for final lengthening in budgerigar vocalizations (Mann, Fitch, Tu, & Hoeschele, 2021), which makes final lengthening an ideal prosodic candidate pattern and budgerigars an ideal model species to conduct behavioral experiments testing the perception and interpretation of prosodic patterns in non-human tetrapods. Such an experiment could look similar to the speech segmentation experiment conducted in my second project, and may use artificial or species-specific vocalizations as target signals, and reaction time measures or touchscreen responses to collect behavioral choices. Using such comparative data, one can gain interesting insights from non-human animal communication research into the phylogenetic origins of speech processing.

3. Contributions to language acquisition and the cultural transmission of languages, and outlook to future research

A central problem in the investigation of language acquisition and language transmission is to identify factors that make linguistic structures easily recognizable and reliably encode a particular meaning. Such structures will be easily acquired, successfully culturally transmitted to future speaker generations and stable over time. My thesis provided several insights into which factors are responsible for the successful encoding of linguistic structure and interpreted them with respect to the cross-linguistic and language-specific occurrence of particular prosodic patterns.

Overall, the thesis showed that cross-linguistically consistent prosodic patterns had a special role in encoding linguistic structure. They were more helpful for speech segmentation, more easily acquirable in a second language and more aesthetically pleasing than language-specific prosodic modifications. This may be linked to their acoustic salience and high occurrence frequency, and therefore to their reliable and unambiguous signaling function.

First, the findings of the speech segmentation project supported the hypothesis that physiologically grounded and thus cross-linguistically consistent cues such as final lengthening (Fletcher, 2010; Tyler & Cutler, 2009) indicate structure unambiguously and are therefore preferentially used for speech segmentation. In contrast, culturally evolved language-specific cues to word stress, which could in theory also act as indicators of wordhood (as in Ordin, Polyanskaya, Laka, & Nespor, 2017), were less relevant in our experiment. Determining and memorizing language-typical prosodic patterns requires computing and constantly updating occurrence frequencies of word stress patterns in the learners' linguistic environment, which may be more cognitively demanding than relying on less ambiguous physiologically grounded cues that have high occurrence frequencies and apply across languages. To test the relationship of using cross-linguistic and language-specific prosodic cues to structure further, future studies could test listeners with other native languages, in particular those with different word stress patterns than German, and explore if they also prefer cross-linguistically occurring boundary cues over language-specific stress cues when segmenting speech. To some extent, such studies exist already (e.g. Ordin et al., 2017; Tyler & Cutler, 2009), but these studies have not yet tested if contrasting cues such as syllable lengthening vs. shortening differed in their contribution to successful speech segmentation. In addition to testing contrasting cues in different languages, future research could expand the thesis' research on contrasting prosodic cues in word-final position to word-initial or word-medial positions.

Secondly, the study on the second language learning of pause patterns points towards a high learnability of acoustically salient cues such as pauses. On the one hand, this may be because pause patterns across dialects and languages have culturally evolved to be cross-linguistically similar and these similar patterns can now be easily transferred from a native to a second language. On the other hand, pause patterns may be easily learnable in a second language irrespective of the native language's pause patterns. In any case, acoustic salience and ease of production of pauses are crucial factors in the encoding of linguistic structure. Using methods inspired by second language acquisition research is novel in the study of the cultural evolution of languages and can be useful to further investigate the learnability and transmissibility of prosodic patterns in various languages that are otherwise difficult to compare. In a next step, our study could be extended to more languages from different language families or to younger, less proficient, learners to draw more wide-reaching and generalizable conclusions about the cross-linguistic and language-specific usage of prosodic patterns as structural cues.

Finally, this thesis includes the first study to address the aesthetic perception of linguistic features as a potential factor in the cultural evolution of languages. Interestingly, also in this study, cross-linguistically occurring prosodic patterns were perceived as aesthetically more appealing than language-specific prosodic patterns. In combination with our speech segmentation experiment, this points towards a potential link between language acquisition and aesthetic appeal. The findings of this thesis provide a foundation for future research in similar directions. First, it is essential to test the aesthetic appeal of prosodic patterns in raters with a different language background than German to see if the preferences established in our study indeed hold cross-linguistically. Secondly, the setup introduced in the thesis also lends itself well to testing the appeal of prosodic patterns other than temporal ones, for example patterns varying in pitch or in a combination of different prosodic modifications. In the future, testing aesthetic appeal may also be extended to other linguistic features such as phonological or grammatical structures. Finally, our study makes the clear prediction that aesthetically pleasing prosodic patterns should be acquired and transmitted more successfully than less appealing ones and that they should thus be diachronically more stable. This could be tested, for example, in an iterated learning study that uses the thesis' findings as a baseline for quantifying aesthetic appeal.

While research in this thesis has mostly focused on prosodic patterns on the word level, future research could be extended to investigations of cross-linguistic and language-specific intonation patterns across larger chunks of speech than words (e.g. intonational phrases or tunes; Crystal, 1969; Hirschberg & Pierrehumbert, 1986; Hirst & Di Cristo, 1998; Ladd, 1978, 2001; Pierrehumbert & Hirschberg, 1990). The production, processing, and perception of larger intonational patterns may be more cognitively demanding than the production, processing, and perception of individual words because larger intonational patterns usually contain more information and per definition occur less frequently than shorter prosodic patterns. Also, in larger prosodic patterns, information has to be retained in memory longer. Although, in general, relations between cross-linguistically occurring and language-specific prosodic patterns during speech segmentation (e.g. Christophe, Peperkamp, Pallier, Block, & Mehler, 2004; Sohail & Johnson, 2016) or aesthetic perception are expected to be similar on the word level and across larger chunks, future research in this direction may provide more thorough insights on the matter.

4. Contributions to cognitive linguistics, historical linguistics and language change, and outlook to future research

One of the crucial questions at the intersection of historical linguistics, psycholinguistics and cultural evolution is whether findings on the synchronic ease of acquisition and transmissibility of particular prosodic patterns can explain actual diachronic language changes. Linked to that, one particularly hot topic in current cognitive linguistic research is the influence of occurrence frequencies of linguistic patterns on cognitive processing (e.g. Divjak, 2019a, 2019b). Do patterns that are frequent, and thus characteristic and reliable cues to linguistic structure get even more frequent over time? Is there evidence for this hypothesis in actual diachronic language data? While occurrence frequencies of lexical or grammatical items have been widely investigated in this regard (e.g. Baayen, Milin, & Ramscar, 2016; Brysbaert, Mandera, & Keuleers, 2018; Bybee, 2007; Diessel, 2007; Ellis, 2002; Lieberman, Michel, Jackson, Tang, & Nowak, 2007; Pagel, Atkinson, & Meade, 2007), research on occurrence frequencies of prosodic patterns has been lacking. By showing in diachronic corpus data that the occurrence frequencies of prosodic patterns indeed influenced cognitive processing and in turn language change, this thesis makes an innovative contribution to historical and cognitive linguistic research.

Since we only tested one particular sound change in one particular language, future studies should perform similar investigations for other sound changes, including sound changes in other languages than English. This could corroborate the hypothesis that attending to statistical regularities between prosodic patterns and lexical or morphological structures results from a general – and potentially evolutionarily old – cognitive bias that acts across dialects and languages through time.

Methodologically, we showed that written historical corpus data can be used not only to answer questions about lexical or grammatical changes, but also about sound changes, which are challenging to extract from written data. This opens new avenues for future diachronic corpus research on sound changes. Ideally, future studies will find a way to largely automatize extracting and coding parameters that were hand-coded in this thesis, such as vowel length, vowel height or morphological structure, using for example machine learning techniques.

Additionally, our study contributes to explaining Middle English Open Syllable Lengthening, which is a widely described but not yet fully explained sound change in the history of English (e.g. Mailhammer, Kruger, & Makiyama, 2015; Minkova & Lefkowitz, 2020). Our findings provide a novel perspective on the implementation pattern of Open Syllable Lengthening, which contributes to solving a long-standing problem in English historical linguistics.

5. Conclusion

Overall, the thesis showed that prosodic patterns were reliable and meaningful cues to linguistic structure. Language learners could use those patterns to segment words and their associated meanings from continuous speech for several reasons: either because they were acoustically salient, aesthetically appealing or frequently occurring. All of these factors made prosodic patterns easily cognitively processable, reliable and unambiguous. The thesis' results indicate that this had a direct influence on language transmission and in turn on diachronic sound changes, which were implemented in a way that increased and stabilized the most characteristic and reliable prosodic patterns.

The thesis points towards a special role of cross-linguistically consistent prosodic patterns for speech segmentation, language learning, language transmission and language evolution.

Especially, the cross-linguistically occurring prosodic pattern of final lengthening was found to be a very reliable and salient cue to linguistic structure. Its cross-linguistic occurrence suggests that final lengthening may be rooted in evolutionarily old physiological and/or cognitive constraints that humans share with other tetrapod species. Thus, in sum, the thesis sheds light on the interplay of biological and cultural evolutionary processes that led to the emergence of the prosodic patterns in present-day languages, past language stages and non-human tetrapod vocalizations, and opens the door to several new lines of research on these topics.

6. References

Baayen, R. H., Milin, P., & Ramscar, M. (2016). Frequency in lexical processing. *Aphasiology*, *30*(11), 1174–1220. https://doi.org/10.1080/02687038.2016.1147767

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The Word Frequency Effect in Word Processing: An Updated Review. *Current Directions in Psychological Science*, *27*(1), 45–50. https://doi.org/10.1177/0963721417727521

Bybee, J. (2007). *Frequency of use and the organization of language*. Oxford: Oxford University Press.

Christophe, A., Peperkamp, S., Pallier, C., Block, E., & Mehler, J. (2004). Phonological phrase boundaries constrain lexical access I. Adult data. *Journal of Memory and Language*, *51*(4), 523–547. https://doi.org/10.1016/j.jml.2004.07.001

Crystal, D. (1969). *Prosodic systems and intonation in English*. Cambridge: Cambridge University Press.

Diessel, H. (2007). Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology*, *25*, 108–127. https://doi.org/10.1016/j.newideapsych.2007.02.002

Divjak, D. (2019a). Counting Occurrences: How Frequency Made Its Way into the Study of Language. *Frequency in Language*, 15–39. https://doi.org/10.1017/9781316084410.002

Divjak, D. (2019b). *Frequency in Language - Memory, Attention and Learning*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316084410

Edwards, J., Beckman, M. E., & Fletcher, J. (1991). The articulatory kinematics of final lengthening. *The Journal of the Acoustical Society of America*, *89*(1), 369–382. https://doi.org/10.1121/1.400674

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition*, *24*, 143–188.

Fletcher, J. (2010). The Prosody of Speech : Timing and Rhythm. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The handbook of phonetic sciences* (2nd ed., pp. 523–602). Hoboken: Wiley-Blackwell.

Friberg, A., & Sundberg, J. (1999). Does music performance allude to locomotion ? A model of final ritardandi derived from measurements of stopping runners. *The Journal of the Acoustical Society of America*, *105*(3), 1469–1484. https://doi.org/10.1121/1.426687

Hirschberg, J., & Pierrehumbert, J. (1986). The intonational structuring of discourse. *Proceedings of the Twenthy-Forth Annual Meeting, Association for Computational Linguistics, Stanford, CA*, 136–144. https://doi.org/10.3115/981131.981152

Hirst, D., & Di Cristo, A. (Eds.). (1998). *Intonation systems: a survey of twenty languages*. Cambridge: Cambridge University Press.

Ladd, D. R. (1978). *The structure of intonational meaning - evidence from English*. Bloomington: Indiana University Press.

Ladd, D. R. (2001). Intonational universals and intonational typology. In M. Haspelmath, E. König, W. Oesterreicher, & W. Raible (Eds.), *Language typology and language universals* (pp. 1380–1390). Berlin: De Gruyter.

Lieberman, E., Michel, J. B., Jackson, J., Tang, T., & Nowak, M. A. (2007). Quantifying the evolutionary dynamics of language. *Nature*, *449*(7163), 713–716. https://doi.org/10.1038/nature06137

Mailhammer, R., Kruger, W. W., & Makiyama, A. (2015). Type frequency influences phonological generalizations: Eliminating stressed open syllables with short vowels in West Germanic. *Journal of Germanic Linguistics*, *27*(3), 205–237. https://doi.org/10.1017/S1470542715000069

Mann, D. C., Fitch, W. T., Tu, H. W., & Hoeschele, M. (2021). Universal principles underlying segmental structures in parrot song and human speech. *Scientific Reports*, *11*(1), 1–14. https://doi.org/10.1038/s41598-020-80340-y

Minkova, D., & Lefkowitz, M. (2020). Middle English Open Syllable Lengthening (MEOSL) or Middle English Compensatory Lengthening (MECL)? *English Language and Linguistics*, (December 2015), 1–26. https://doi.org/10.1017/S1360674319000522

Morton, E. S. (Ed.). (2017). *Animal vocal communication - assessment and management roles*. Cambrdige: Cambridge University Press.

Ordin, M., Polyanskaya, L., Laka, I., & Nespor, M. (2017). Cross-linguistic differences in the use of durational cues for the segmentation of a novel language. *Memory & Cognition*, *45*, 863–876. https://doi.org/10.3758/s13421-017-0700-9

Pagel, M., Atkinson, Q. D., & Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature*, *449*, 717–720. https://doi.org/10.1038/nature06176

Pierrehumbert, J., & Hirschberg, J. (1990). The Meaning of Intonational Contours in the Interpretation of Discourse. In P. R. Cohen, J. Morgan, & M. E. Pollack (Eds.), *Intentions in communication*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/3839.003.0016

Sohail, J., & Johnson, E. K. (2016). How Transitional Probabilities and the Edge Effect Contribute to Listeners ' Phonological Bootstrapping Success. *Language Learning and Development*, *12*(2), 105–115. https://doi.org/10.1080/15475441.2015.1073153

Toro, J. M., & Crespo-Bojorque, P. (2021). Arc-shaped pitch contours facilitate item recognition in non-human animals. *Cognition*, *213*(July 2020). https://doi.org/10.1016/j.cognition.2021.104614

Tyler, M. D., & Cutler, A. (2009). Cross-language differences in cue use for speech segmentation. *Journal of the Acoustical Society of America*, *126*(1), 367–376. https://doi.org/10.1121/1.3129127

Abstract

Spoken language is our main communication system, allowing us to share our thoughts and ideas with others. There is a large variety of different dialects and languages, and human language is strikingly different from non-human tetrapod vocal communication systems. Still, when taking a closer look, there are also many linguistic features that are shared across dialects, languages and species. This thesis addresses the question of which factors determine these differences and similarities.

Current research explains linguistic diversity by how easily particular linguistic features are acquired and transmitted to future generations of speakers. Differences in the learnability and transmissibility of speech characteristics in individual communities may cause dialects and languages to change and in turn to become distinct. In contrast, linguistic features that are determined by evolutionarily old physiological or basic cognitive constraints, are likely to be similar across languages and species.

One particularly interesting characteristic of spoken language, which is relatively similar across languages and species, but still differs in its details, is prosody, the acoustic modulation of speech. Prosody subsumes vocal features such as variations in pitch, duration, and speech pauses.

This thesis investigates how prosodic patterns contribute to encoding structure in vocal signals, thus facilitate language learning, and in turn influence language change and language evolution. Reviewing prosodic patterns in human and non-human tetrapod vocalizations, chapter 1 suggests that prosodic patterns that occur cross-linguistically and thus possibly have evolutionarily old origins may be processed similarly by humans and non-human tetrapods. Chapter 2 investigates the effects of various prosodic patterns on speech segmentation, finding that acoustically salient and cross-linguistically consistent cues such as pauses and final lengthening are more relevant than language-specific stress patterns for recognizing and acquiring words in continuous speech. Chapter 3 introduces differences in the aesthetic perception of prosodic patterns as potential factors that may determine the learnability and transmissibility of language. Chapter 4 reveals that native and non-native speakers of English realize speech pauses very similarly, suggesting that general physiological and cognitive processes lead to a high learnability of pauses across languages. Finally, chapter 5 shows that

the learning biases addressed in the previous chapters manifest in actual diachronic language change by demonstrating that the prosodic shapes of newly emerging words in the history of English assimilated to the prosodic shapes of highly frequent, already existing words. Thus, overall, this thesis helps to elucidate how biological and cultural pressures interact in language learning, transmission, change and evolution.

Zusammenfassung

Sprache ist das wichtigste Kommunikationssystem der Menschen und erlaubt ihnen, ihre Gedanken und Ideen mit anderen zu teilen. Sprache ist vielfältig: es gibt eine große Fülle an verschiedenen Sprachen und die menschliche Sprache unterscheidet sich deutlich von tierischen vokalen Kommunikationssystemen. Dennoch gibt es einige sprachliche Merkmale, die alle menschlichen Sprachen und sogar tierische Kommunikationssysteme gemeinsam haben. Diese Dissertation untersucht, welche Faktoren diese Unterschiede und Gemeinsamkeiten bestimmen.

Linguistische Diversität wird in der aktuellen Forschung oft dadurch erklärt, wie leicht bestimmte sprachliche Merkmale gelernt und an nachfolgende Generationen weitergegeben werden können. Unterschiede in der Lernbarkeit und Weitergebbarkeit von sprachlichen Merkmalen in verschiedenen Sprachgemeinschaften können dazu führen, dass sich Sprachen verändern und in weiterer Folge immer mehr unterscheiden. Im Gegensatz dazu ähneln sich linguistische Merkmale in verschiedenen Sprachen und Tiervokalisationen oft, wenn sie durch evolutionär alte und demnach grundlegende physiologische und kognitive Eigenschaften beeinflusst werden.

Ein besonders interessanter Aspekt der vokalen Kommunikation, welcher sich in menschlichen Sprachen und tierischen vokalen Kommunikationssystemen ähnelt, ist Prosodie (Sprachmelodie). Unter Prosodie werden vokale Modifikationen wie zum Beispiel Variationen in der Silbenlänge, in der Tonhöhe oder in Sprachpausen zusammengefasst.

Diese Dissertation erforscht, wie Muster in der Sprachmelodie dazu beitragen, Strukturen in Lautäußerungen zu schaffen, auf diese Weise das Sprachenlernen erleichtern und in weiterer Folge Sprachwandel und Sprachevolution beeinflussen. Kapitel 1 stellt sprachmelodische Muster in menschlichen und tierischen Vokalisationen gegenüber und erläutert, dass sprachenübergreifend vorkommende sprachmelodische Muster vermutlich evolutionär alte Ursprünge haben und von allen Landwirbeltieren ähnlich verarbeitet werden. Kapitel 2 untersucht die Effekte von unterschiedlichen sprachmelodischen Mustern auf das Wortlernen. Akustisch prominente und sprachübergreifend vorkommende sprachmelodische Muster, wie zum Beispiel die Verlängerung von Silben am Phrasenende, waren relevanter für das Erkennen von Wörtern in einem kontinuierlichen Sprachfluss als sprachspezifische Betonungsmuster.

Kapitel 3 fand, dass verschiedene sprachmelodische Muster als unterschiedlich schön wahrgenommen werden. Dies legt nahe, dass die ästhetische Wahrnehmung von linguistischen Merkmalen beeinflussen kann, wie leicht diese Merkmale gelernt und an andere Personen weitergegeben werden können. Kapitel 4 fand, dass Personen, die Englisch als Zweitsprache sprechen, ihre Sprachpausen sehr ähnlich machten wie Personen, die Englisch als Erstsprache sprechen. Dies weist darauf hin, dass allgemeine physiologische und kognitive Prozesse zu einer hohen Lernbarkeit von Pausen unabhängig der Sprache führen. Kapitel 5 zeigt, dass sich die Unterschiede in der Lernbarkeit von verschiedenen sprachmelodischen Merkmalen, die in den vorherigen Kapiteln experimentell untersucht wurden, auch in tatsächlichen Sprachwandelprozessen zeigen. In der englischen Sprachgeschichte passten sich nämlich die sprachmelodischen Formen von neu entstehenden Wörtern an die Formen von häufig vorkommenden und dadurch leicht lernbaren schon existierenden Wörtern an.

Zusammenfassend informiert die Dissertation also darüber, wie biologische und kulturelle Faktoren beim Sprachenlernen, der Weitergabe von Sprache, dem Sprachwandel und der Sprachevolution zusammenspielen.