Enhancement effects of frequency: An explanation from the perspective of Discriminative Learning

DISSERTATION

zur

Erlangung des akademischen Grades

Doktor der Philosophie

in der Philosophischen Fakultät

der Eberhard Karls Universität Tübingen

vorgelegt von

SAITO, MOTOKI

aus

Tokyo, Japan

Tübingen

2024

Gedruckt mit Genehmigung der Philosophischen Fakultät
der Eberhard Karls Universität Tübingen.

Dekanin:	Prof. Dr. Angelika Zirker
Hauptberichterstatter:	Prof. Dr. R. Harald Baayen
Mitberichterstatter:	Dr. Fabian Tomaschek
Mitberichterstatter:	Prof. Dr. Ruben van de Vijver
Tag der mündlichen Prüfung:	17. November, 2023.

Universitätsbibliothek, TOBIAS-lib
Eberhard Karls Universität Tübingen
Tübingen

To Seira, and to Hyugo

Acknowledgements

This dissertation would not have been possible without help from a lot of people. First of all, I would like to show my deepest gratitude to my supervisor, Harald Baayen. He literally saved my life and my career. I have been interested in speech processing including comprehension and production both all the time since my bachelor study. However, I was struggling to find a suitable place for me and for my interest in speech processing in Japan. He offered me an opportunity to work in psycholinguistics. He accepted me and taught me countlessly many topics from many different perspectives. Throughout the years, by meetings and casual conversations, he always guided me in the right direction. He sometimes helped me expand my knowledge about theories, frameworks, models, and the literature I had not known. Other times, he helped me improve my practical skills by showing me what kinds of analyses are possible and necessary, how to perform analyses more efficiently, and so on, among others. He also supported me emotionally as well. He encouraged me always with his positive and kind attitude and personality. Without him, I would not have been able to be here, I would not have been able to acquire these knowledge and skills I have now, and I do not know even if I would have been pursuing this career as a researcher. I am really greatly thankful for you, Harald.

Furthermore, I would really like to show my gratitude to my second supervisor, Fabian Tomaschek, from my heart. He always put up with my stubbornness, spending a lot of time for me. With his wide range of knowledge and skills, he also guided me to the literature and the papers I had not known yet and helped

me improve my analysis skills. He was always not only a good mentor for me, but also really a good friend of me. He always reminded me of the importance of my own private life with my family. He also helped my life in many respects in Germany. My life here in Tübingen is greatly indebted to his kindness, generosity, and warmhearted personality.

I am also obliged to Ingo Plag and his research team members in the DFG research project "FOR2373 Spoken Morphology". The project's colloquium gave me a lot of insights constantly for my own project and have always been a continuous source of inspiration. When I had my own presentation in the colloquium, everybody took time generously and helped me develop my project with a lot of advice from many different perspectives, which I could not have been able to come up with or even notice by myself.

Moreover, I would like to thank every member of the Quantitative Linguistics research group. Tineke drastically reduced the time I would have had to spend for administrative issues without her. When I had to do something for administrative purposes and when I had questions related to administrative paperwork, I simply went to and asked her about them. She always helped me in a very kind and cheerful manner. Tino maintained the servers for the group in the first place, without which I would not have been able to do most of the analyses in this dissertation. Moreover, he helped me every time very kindly when I had technical problems with my computers and my scripts. Furthermore, he helped me translate the discussion chapter of this thesis in German. June provided me the "second home" for me in the office. She also learned a lot about the ultrasound system we have in our group and summarized important information. Thanks to her, I could learn the setup and the usage of the ultrasound system fairly quickly. Peter helped me many times as well when I had problems and questions in my script and analysis. Although he was very busy in those days, which I know now, he always helped me very kindly. Michael helped me to get into the world of discriminative learning models. I learned a lot of the basics and aspects of discriminative learning from

him. In addition, I would like to thank the entire group for such an inspiring and, at the same time, such a cozy research environment. Elnaz, Inna, Jessie, Karen, Karlina, Kun, Maja, Maria, Paula, Sean, Yang-Yi, Yu-Xing, Yu-Ying, I thank you all for such a nice time.

Finally, I would really like to thank my family. My parents have always been supportive for my personal and academic dreams. They were always my role models. Without them, I would not have been able to pursue this career in the first place. And, of course, my partner, Seira. She gave up her work, her friends, and her own family in Japan all for me, and came to Germany. She encountered a lot of difficulties since then. To say nothing of language problems, cultural differences between Asia and Europe were so huge that she had to struggle for such a long time. Nevertheless, she kept making effort every day to integrate herself in Germany and to make our life work here in Germany. On the top of these, to which I already cannot thank enough, she carried our son, Hyugo, for 10 months in her body and brought him to the world, going through such strong pain and hardships, which I cannot even imagine. Even after Hyugo was born, she mainly took care of Hyugo all for me, so that I could focus on my own work. Clearly, this dissertation would have been impossible without her constant support. I really appreciate her support. I do really not know how I can thank you enough, Seira. And, Hyugo, you and your smile always give me happiness, pleasure, and strength to work through any hardship. Please forgive me that I could not have so much time to play with you, Hyugo. Let us play together more often, and let us go for walking, hiking, and exploring in the nature more often, I promise. I am such a lucky person to have you two, Seira, Hyugo. I am so looking forward to our life unfolding in front of us. I love you.

Abstract

Frequency and frequency-based measures such as probability are widely accepted to have extensive effects on speech production. On the one hand, high frequency units (e.g., words) have been found to be articulated with shorter duration and centralized tongue positions, indicating phonetic reduction. On the other hand, high frequency, and therefore more predictable, units have been found to be articulated with longer duration and clearer articulations, indicating phonetic enhancement.

This dissertation provides a possible account for these seemingly-contradictory effects of frequency from the perspective of semantics. To this end, I first replicated one previous study, using ultrasound, which reported phonetic enhancement effects of frequency based on tongue position data recorded by electromagnetic articulography (EMA). In addition, I developed a new methodology of analyzing ultrasound images, in which not only tongue surface positions but the whole ultrasound images can be included for analysis. Using a different recording technique and a different analysis technique, effects of phonetic enhancement of word frequency were replicated.

Subsequently, I collected a set of words that shared the same rime structure, namely the stem vowel [a(ː)], at most one intervening segment, and the word-final plosive [t], from a spontaneous speech corpus. Critically, these words differed in their morphological status, having either one or no morphological boundary between the stem vowel and the word-final [t], namely inflected and non-inflected words with the same rime structure. For these words, an increase in frequency was observed to be associated with higher tongue positions (indicating phonetically

reduced realizations) of the stem vowel, while phonetic reduction was extensively attenuated for inflected words. These results suggest that seemingly-contradictory effects of frequency may in fact be due to different morphological statuses of the items being investigated.

Finally, a possible source of this "morphological" modulation of the frequency effect was investigated from the perspective of semantics. The amount of semantic support for the word-final triphone (i.e., `SemSupSuffix`), which contains the suffix in the middle, was calculated from the Discriminative Lexicon Model (DLM), which was trained to discriminate all the German words available in the CELEX database. `SemSupSuffix` was found to be associated most clearly with the word's inflectional status in such a way that higher `SemSupSuffix` was more likely for inflected words. Furthermore, `SemSupSuffix` showed a better performance in predicting tongue positions during the stem vowel in a statistical model, compared to inflectional status as a dichotomous factor variable. Moreover, the model with `SemSupSuffix` predicted that high frequency was associated with phonetic reduction when `SemSupSuffix` was low, which corresponded to non-inflected words, and that it was associated with phonetic enhancement when `SemSupSuffix` was high, which corresponded to inflected words. These results clarify that the observed interaction of frequency by morphological status can be explained without resorting to morphological concepts such as morphemes, because the DLM, from which `SemSupSuffix` was derived, does not make use of such high-level constructs, but uses lower-level sublexical form features instead.

In summary, this dissertation provides an explanation for why effects of frequency are different for inflected and non-inflected words, without requiring the theoretical constructs of morpheme or exponent. This explanation builds on the strong support that a low-level sublexical unit (the triphone straddling the suffix /t/) receives from a words' semantics. Instead of calling upon a putative morpheme boundary that would be part of a words' form representation, this explanation points to the importance of the different semantics that are realized in the word

and the differential support that the critical triphone receives from the semantics.

Contents

List of Figures

List of Tables

Chapter 1

Background: Speech processing from semantics to phonetics

Abstract: This chapter provides an overview of models of speech production that start from conceptual and semantic levels and proceed all the way down to phonetic realizations. Early models (Fromkin, 1971; Garrett, 1984; Levelt & Wheeldon, 1994) were verbal models that were highly modular and serial with architectures consisting of several hierarchically organized distinct levels or modules. Individual models of speech production differ in the extent to which modularity and seriality are imposed. Another influential model (e.g., Dell, 1986) implemented interactive activation network with limited seriality. A more recent computational model linking meaning to form (Baayen et al., 2019) works with a simple network with no intermediate layers between high dimensional representations of meanings and high dimensional representations of forms.

1.1 Classical perspectives on speech production

An important part of the speech production process involves selecting those words that properly realize the meanings that the speaker wants to express. It is uncontroversial that the speech production process starts off with concepts and meanings and culminates in articulation. However, it varies from model to model what intermediate processing stages and how many processing stages are involved. Most of the earliest models of speech production work with intermediate levels that follow standard components of the grammar as laid out in formal theories within the framework of generative grammar, with Chomsky (1965) being particularly influential.

1.1.1 Fromkin's model

In early days, speech production models were constructed based on observations of speech errors, the assumption being that speech errors are revealing about the internal structure of the cognitive system that drives the speech production process. For example, if swapping of two phonemes is observed (e.g., *keep a tape* → *teep a cape*), the production system was assumed to operate on phonological units. If, on the other hand, a certain unit is not involved in speech errors, the unit was assumed to be irrelevant to the speech production system. The internal structure of affricates, for instance, are not split up in speech errors (e.g., *pinch hit* → *pinch hitch*, but not [pɪnt hɪʃ]) (Fromkin, 1971), indicating that affricates are phones for English speakers rather than sequences of phones.

Based on detailed analysis of speech errors, which were collected by the author herself throughout three years of her academic and personal life, Fromkin (1971) proposed six stages of processing. The first stage is the generation of an intended meaning. Little detail was provided as to the structure within this stage and its connection to the next stage, the syntactic stage. In this next stage, a syntactic structure is generated. This syntactic structure has slots for words. These slots are

specified for semantic and syntactic features. This stage was supposed to explain speech errors that involve switching words in the same syntactic word class (e.g., nouns exchanged with other nouns).

After the syntactic structure is created, at stage three, an intonation contour is assigned to the syntactic structure, distinguishing, for instance, between declarative and interrogative sentences. Such an intonation contour of an entire sentence was posited to explain the preservation of sentential stress positions. For example, according to Fromkin (1971, p. 42), the second word holds the primary sentential stress when *How bád things were* is mis-uttered for *How thíngs bad were*.

At the next stage, the lexicon comes into play. Definitions of the lexicon differ from model to model. In her model (Fromkin, 1971), the lexicon was assumed to consist of two parts: a semantic section and a phonological section[1]. In the semantic section, words were assumed to be specified with semantic features, further grouped into syntactic categories such as nouns and verbs. These entries in the semantic section have a pointer to their corresponding entry in the other part of the lexicon, the phonological section, in which segmental information of words is specified.

According to Fromkin (1971), look-up in the lexicon consists of two steps. Words are first looked up in the semantic section based on the semantic and syntactic features assigned to the word slots in the syntactic structure. Speech errors involving words with similar meanings (e.g., *like* for *hate*) can occur during this selection stage, due to sharing many semantic features. Second, once an entry in the semantic section has been accessed, look-up reads the pointer in the entry and proceeds to the specific entry indicated by the pointer in the phonological section of the lexicon. Words' phonological segments are specified in the phonological section. Since words sharing similar phonological segments are assumed to be located in the vicinity in this section of the lexicon, word-switching based on phonological similarity (e.g., *pressure* for *present*) is argued to occur at this stage of identifying

[1]These were called "the semantic class sub-section" and "the over-all vocabulary" (Fromkin, 1971).

an entry in the phonological section. Access to an entry in the phonological section produces a string of phonological segments, each of which is specified with respect to its syllabic and segmental position. As a consequence, Fromkin (1971) claimed, misordering of segments within/across syllables occurs during this process of looking up the target words and citing their phonological segments in a string.

Subsequently, these phonological segments are brought together in syllables and, where necessary, modified by morphophonemic constraints. Fromkin (1971) claimed that segmental speech errors involving allomorphs (e.g., /s/ or /z/ for the plural suffix) should occur before this stage, but without providing much detail. At the final stage, the phonemes in the syllables are converted into actual neuro-motor commands driving articulation.

1.1.2 Garrett's model

The speech production model by Fromkin (1971) was later refined by Garrett (1984, 1988). His model posits five stages: the message level, the functional level, the positional level, the phonetic level, and the articulatory level. In each of the levels, representations from the level immediately above were received as input, certain types of processes operate on the input representations, and different types of representations are produced as output (Table 1.1). The positional level and the phonetic level can be combined and treated as jointly constituting the positional level. In addition, the functional, positional, and phonetic levels jointly constitute the sentence level.

At the message level, general concepts are taken as input, inferential processes operate on these general concepts, and message-level representations are produced as output. Garrett (1984) suggested that a conceptual syntax operates on general concepts and builds a more complex representation out of them in a compositional way. However, this level was not regarded as "linguistic", and therefore not much detail was provided.

Table 1.1: The five levels and their processes, input representations, and output representations. For example, the functional level receives message-level representations and returns functional-level representations.

		Representation	
Level	Process	Input	Output
Message	Inferential	General concepts	Message-level
Functional	Logical/syntactic	Message-level	Functional-level
Positional	Syntactic/phonological	Functional-level	Positional-level
Phonetic	Regular phonological	Positional-level	Phonetic-level
Articulatory	Motor coding	Phonetic-level	Articulatory

At the functional level, message-level representations are taken as input, logical/syntactic processes operate on them, and functional-level representations are produced as output. The processes at this level consist of lexical selection (retrieval of lexical items), construction of a frame (structure), and assignment of the retrieved items to the slots of the frame. In these processes, meaning-based word substitution errors and whole-word exchange errors can occur such as (1) and (2) below respectively.

(1) He rode his bike to school *tomorrow*. (yesterday)

(2) Cat, it's too *hungry* for you to be *early*.

Meaning-based word substitutions are based on semantic similarity. In Garrett's model, meaning-based word substitution errors were assumed to occur in the process of retrieving words based on words' meanings, namely message-level representations. In contrast, whole-word exchanges do not always have semantic similarity between the interacting words. However, whole words are exchanged, not only portions of words, and whole-word exchanges always involve words in the same syntactic category. Based on these observations, whole-word exchanges were assumed to occur in the process of assigning the retrieved words into the corresponding slots in the frame (structure).

The processes at the next level, namely the positional level processes, pro-

ceed in a similar manner as the functional level processes. These processes are retrieval of segmental information for each word, construction of a phonological structure, and assignment of segments to the phonological structure. The input of the positional level is the output of the functional level, namely functional-level representations, based on which segments of each lexical item are retrieved. While segments are retrieved, a surface phrasal geometry (i.e., phonological structure) is constructed, that is, a tree-like phonological structure with slots for segments. The retrieved segmental information is then assigned to these slots. Two kinds of speech errors can occur in this level in a similar way as in the functional level. The retrieval of segmental information can cause word-substitution errors based on form (sound) similarity, and the assignment of segmental information to the phonological structure can lead to sound-exchanges, which are also based on form (sound) similarity. The examples (3) and (4) below correspond to these two kinds of speech errors respectively.

(3) It looks as if you're making considerable *process*. (progress)

(4) I was just gonna ***r****ock* on the ***n****ong* door.

Garrett's model differs from Fromkin's model in that it makes a distinction between the treatments of major-class items (i.e., content words) such as nouns and minor-class items (i.e., function words) including bound morphemes such as inflectional suffixes. This distinction was based on the observations that minor-class items such as suffixes can be shifted to another position in the phrase but not substituted or exchanged for another word or morpheme. In the following example (5), the inflectional suffix *-s* is stranded in its original position, while the root the suffix is attached to (i.e., *pay*) is exchanged for another word in the sentence (i.e., *wait*). In contrast, the following example (6) shows that a bound morpheme can be shifted within a phrase.

(5) It wait*s* to pay. (pays to wait)

(6) I'd forgot__ about*en* that.

Based on these kinds of speech errors, Garrett's model assumed that minor-class elements are associated with the phonological frame. Retrieval failure is the source of substitution errors, such as examples (1 and 3). Assignment of retrieved items to wrong slots of a frame is the source of exchange errors, such as example (2 and 4). In contrast to minor-class items, major-class items such as nouns were assumed to always go through all the processes, namely being retrieved and assigned to the functional frame, followed by their segmental information being retrieved and assigned to the positional (phonological) frame. As a consequence, all kinds of speech errors can be observed for major-class items. Minor-class elements can only be shifted within a phrase. This observation was explained by postulating "error processes in general" (Garrett, 1984, p. 180) that can occur after the positional-level representations are completed, namely at the phonetic level.

The output of the positional level, namely positional-level representations, is then modified by regular phonological processes, and all the segments in the positional-level representation are then concatenated into a string of phonemes. This stage is called the phonetic level. Misplacement of segments can occur in this level at the moment that segments are made into a string, due to what Garrett (1984) called error processes in general. Moreover, this level was assumed to explain phonological accommodation, the phenomenon that an appropriate allomorph is selected and phonologically realized according to its new environment created by a speech error, such as below:

(7) *a* money's aunt (...). (an aunt's money)

In this example, the indefinite article is realized as *a* /ə/ (not *an* /ən/), so that it conforms to its new environment, namely before /m/. This modification was assumed to be brought about by regular phonological processes, which operate in this phonetic level.

The output string of the phonetic level, which is namely the end product of the sentence level (i.e., the functional, positional, and phonetic levels), is then sent to the articulatory level. At this level, the phonetically rich sentence representation

is translated into an articulatory structure that provides motor instructions for the articulators.

1.1.3 Dell's model

In contrast to the serial feedforward modular nature of the models by Fromkin (e.g., Fromkin, 1971) and Garrett (e.g., Garrett, 1984), the model by Dell and his colleagues (Dell, 1986, 1988, 1990; Dell et al., 2007; Dell et al., 1997; Foygel & Dell, 2000; Kittredge et al., 2008) has a network structure, and selection of linguistic units (e.g., word) was assumed to be based on activation levels of units in the lexical network.

This model has four levels: the conceptual level, the syntactic level, the morphological level, and the phonological level[2]. The model is worked out in detail for the last three levels (Dell, 1986). Each level has a "rule" component in addition to a "lexicon" component (i.e., a lexical network). In the rule component (i.e., the tactic frames), a tree-like structure with slots is constructed (e.g., a syntactic frame). Each slot is specified with respect to the appropriate categories at a given level. These categories are syntactic classes such as nouns for the syntactic level, morpho-syntactic classes such as the noun-stem for the morphological level, and positions within a syllable such as nucleus for the phonological level. According to these category specifications, the lexical network associated with a given level (e.g., the lexical network of the syntax level) is searched for the node of the specified category with the highest activation level at a given point in time. Nodes in the network are words at the syntactic level, morphemes at the morphological level, and phonemes at the phonological level. Nodes are not fully connected between adjacent layers, they are connected only to pertinent nodes. After being selected, a node gets associated with a particular slot in the frame.

[2]Dell (1986) also proposed a phonological encoding model, which contained morpheme nodes, syllable nodes, rime nodes, phoneme nodes, and feature nodes. This phonological encoding model can be understood to correspond to the phonological level. However, this phonological encoding model was evaluated separately from the entire model (which contained concepts, words/lemmas, morphemes, and phonemes). In what follows, the focus is on this main model

Processing in the lexicon is governed by spreading activation. At each time step, a node sends a fixed proportion of its activation to its neighbor nodes that are directly connected to the node. At a destination node, arriving activations are summed and added to the current activation level of the node. At the same time, activation levels of nodes are assumed to decay at a fixed rate at each time step to keep activation levels down. Importantly, spreading of activation is assumed to be two-way, not only downwards (i.e., from the conceptual level, through the syntactic and morphological levels to the phonological level) but also upwards (i.e., from the phonological level, through the morphological and syntactic levels to the conceptual level). This characteristic enables nodes at lower levels to affect nodes at higher levels. As a consequence, lower levels can change activations at higher levels, which in turn will affect how higher levels affect lower levels.

Speaking is assumed to begin with increases in activation levels of pertinent conceptual nodes. Conceptual nodes with positive activation levels activate the lemma nodes with which they are connected. The lemma nodes pass on activation to lower levels in the network and subsequently receive activation back from these lower levels. In this process, one lemma will reach a higher activation than the other lemmas, which makes it eligible for insertion into a syntactic tree, under the condition that this lemma is of a word category that matches the slot in the syntactic tree. After selection and lexical insertion, the lemma node is deactivated. A similar procedure is used to select morphemes for insertion into morpheme trees, and to select phonemes for insertion into syllable trees.

For example, to produce the sentence *some swimmers sink*, the syntactic frame shown in Figure 1.1 is constructed. Note that the end nodes of the frame specify a syntactic category for each slot (e.g., Q = Quantifier). Based on this frame, the model first looks for the node of the quantifier with the highest activation level in the lexical network. The specification of the currently appropriate category ensures that the right node belonging to the correct category gets selected. In the current example, *some*, *swimmer*, *sink*, and PLURAL are expected to be activated strongly,

since they should all be activated by the concept that the speaker intends to convey (i.e., the concept of *some swimmers sink*). However, the model only checks the node of the quantifier at the moment in this example, and therefore another highly activated node such as *swimmer* will not be selected.

As a consequence, unless some other quantifiers get activated more strongly than the correct quantifier for some unexpected reasons, the correct lemma node *some* should hold the highest activation level at the time of the node-checking, and the model finds this lemma node. Once the node gets selected, the node gets tagged and associated with the current slot (e.g., "(1)" as shown in the figure), and the model looks for the next appropriate node for the next slot, which is a noun in this example. Then, the selected noun, which should be *swimmer*, will be tagged (e.g., "(2)"). In Figure 1.1, the item being searched for at the moment is marked as "?".

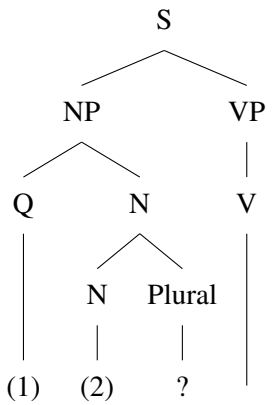


Figure 1.1: The syntactic frame that would be constructed to produce *some swimmers sink* at the moment the model is searching for an inflectional suffix (Dell, 1986). (Q=Quantifier, N=Noun, V=Verb, Plural=a bound morpheme for plurality)

In parallel to the syntactic level, the morphological frame is constructed in the "rule" section at the morphological level (Figure 1.2). In the lexical network, the *current* node is determined. The current node is the node tagged first at the immediately upper level, which is *some* at the syntactic level for this example. The current node receives the initial boost of activation level, which Dell (1986) called

signaling activation, and spreads its activation to its neighbor nodes. In the current example, the activation spreading from the current node *some* should activate the quantifier stem *some* the most strongly, leading to the node being selected for the morphological frame.

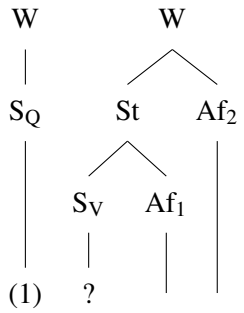


Figure 1.2: The morphological frame that would be constructed to produce *some swimmers (sink)* at the moment the model is searching for a verb stem. (W=Word, S_Q=Quantifier stem, St=Stem, Af=Affix, S_V=Verb stem)

Once an appropriate node gets selected and tagged, the current node is moved to the next slot of the frame, which is "N", namely *swimmer* in the lexical network, in this example. As the current node, the lemma node *swimmer* gets activated to a certain extent (i.e., signaling activation) and spreads its activation to its neighbor nodes.

Note that *swimmer* is a derived form at the syntactic level. Dell (1986) assumed that the derived word has its own node at the syntactic level (e.g., *swimmer*) and it is decomposed into its component morphemes at the morphological level (e.g., *swim* and *er*). In contrast, the inflected word was assumed to be decomposed already at the syntactic level, not only at the morphological level. Therefore, in the present example *(some) swimmers (sink)*, the nodes *swimmer* and PLURAL are activated separately for the word *swimmers* at the syntactic level, and they are decomposed furthermore into *swim*, *-er*, and *-s* at the morphological level. Dell (1986) assumed this distinction between inflection and derivation, mainly following the standard linguistic account about inflection and derivation (Chomsky, 1965, 1981) and some

psycholinguistic studies (Garrett, 1984). In addition, Dell (1986) suggested that his model can explain why suffix shift errors such as (8) tend to occur for inflectional suffixes, rather than derivational suffixes. For an inflectional word, its stem and its suffix are separated at the syntactic level, and each of them sends activation to its corresponding morpheme node separately. As a consequence, the correct stem node may not be activated the most when the suffix node is activated. If the suffix node gets activated, while another (wrong) word (e.g., *get*) is still activated, then the suffix may get attached to the wrong word. In contrast, a derived word has a single lemma at the syntactic level. Its corresponding stem and suffix at the morphological level are both connected with and therefore receive activation at the same time from the lemma node. Since the stem and the suffix nodes receive activation at the same time, the mistiming of constructing the stem-suffix combination does not happen for a derived word.

(8) gets it (for *get its*)

Following the selection of the lemma *swimmer* as the current node, the morpheme nodes *swim* and *-er* get activated. Since the next slot to be filled in the morphological frame is a verb stem, only the morpheme nodes of the verb stem are searched to determine which one of them retains the highest activation. Assuming nothing irregular happens, the morpheme node *swim* should then be selected for the slot. The search then continues to the next slot, which is a derivational suffix in this case.

This cycle from the selection of the current node to the selection of the appropriate node for the slot in the frame works in the same way also for the phonological level. At the phonological level, each node in the lexical network represents a phoneme with its phonological category specified such as onset consonant, nucleus vowel, and coda consonant. For the current example, a phonological frame such as shown in Figure 1.3 would be constructed for *some*.

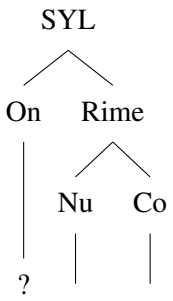


Figure 1.3: The phonological frame for *some* at the moment the model is searching for the onset consonant (Dell, 1986). (SYL=Syllable, On=Onset, Nu=Nucleus, Co=Coda)

In this model, speech errors are explained mainly in terms of the mis-selection of the node due to particular activation patterns. For example, non-contextual errors based on semantic similarity such as (9) are assumed to occur due to the shared lemma. In the case below, SINK is connected to *sink*, and *sink* is connected to DROWN. Since DROWN has its own corresponding lemma *drown*, it is possible that activation originating in SINK travels through *sink* and DROWN, activating the wrong lemma *drown* in the end. If the model looks for the most activated node in the lexical network in order to fill in the slot of the verb in the syntactic frame when the wrong lemma *drown* happens to have a higher activation level than the correct lemma *sink*, the word substitution from *sink* to *drown* might occur.

(9) Some swimmers drown. (for *Some swimmers sink*)

Such a "mis-activation" can also occur because the processing at lower levels necessarily follow the processing at higher levels. In other words, phonological processing (e.g., selection of phonemes) begins once at least one morpheme node is selected and tagged for its corresponding slot in the morphological frame, and morphological processing begins once at least one lemma is selected. As a consequence, it is quite possible that the morphological level already works on the second word, while the phonological level is still working on the first word. This can lead to anticipatory errors such as (10) below:

(10) sim swimmers /sɪm swɪmɚz/ (for *some swimmers* /sʌm swɪmɚz/)

In this example, the morphological level may already be working on the noun stem *swim*, namely looking for the morpheme node *swim* to fill in the slot of the noun stem in the morphological frame. At this moment, the morpheme node *swim* should have a very high activation level, at least higher than the other morpheme nodes of the noun stem. As a consequence, the morpheme node *swim* spreads a considerable amount of activation to its constituent phonemes, including /ɪ/. Meanwhile, the phonological level may still be trying to find the nucleus vowel for *some*, where /ʌ/ should be found. However, since /ɪ/ may also be highly activated due to the activation spread from *swim*, it is possible that /ɪ/ happens to have a higher activation level than /ʌ/, leading to the wrong selection of /ɪ/ as the nucleus vowel of *some*, which results in /sɪm/ in place of /sʌm/.

The model assumes that the activation level of a node that has already been selected gets reduced to zero. However, its neighboring nodes usually retain a relatively high activation level, since the node that has been selected and deactivated should already have sent some activation to its neighboring nodes before the deactivation. As a consequence, the very node that has been deactivated will receive a considerable amount of activation from its neighboring nodes after the deactivation, leading to a high activation level again. Dell (1986) called this "rebounding". This "rebounding" mechanism was claimed to explain perseveration errors such as (11) below:

(11) some swummers /sʌm swʌmɚz/ (for *some swimmers* /sʌm swɪmɚz/)

After the selection of the phoneme node /ʌ/, the node gets deactivated. However, due to the "rebounding" mechanism, the activation level of the node rebounds back to being high. When the rebounded activation level happens to be higher than the activation level of the phoneme node /ɪ/ at the moment looking for the nucleus vowel for *swim*, the perseveration error in (11) occurs.

Exchange errors such as shown in (12) are explained by the combination of the

mechanisms responsible for anticipatory errors and perseveration errors.

(12) sim swummers /sɪm swʌmɚz/ (for *some swimmers* /sʌm swɪmɚz/)

When the anticipatory error occurs, the wrongly-selected node (i.e., /ɪ/) gets deactivated. As a consequence, the other phoneme node, which would have been correct for the last slot but wrong for the current slot, namely /ʌ/ in this example, is possible to have a higher activation level than the correct one (i.e., /ɪ/). This can lead to selecting /ʌ/ for the nucleus vowel of *swim*, resulting in the exchange error shown in (12).

This model by Dell and his colleagues (e.g., Dell, 1986) is different from the preceding classic models (Fromkin, 1971; Garrett, 1984) in that the selection of linguistic units (e.g., word) is completely based on activation spreading in the lexical network and that it allows lower levels to affect higher levels by spreading activation. In addition, Dell's model is more flexible which enables it to capture intermediate cases of speech errors. For example, suppose the speaker said *Let's stop* instead of *Let's start*. This speech error can be understood as an instance of the word exchange based on semantic similarity. At the same time, since *stop* and *start* share the same consonant cluster at the syllable onset, the speech error can also be classified as a speech error based on phonological similarity. Nevertheless, in serial feedforward modular models (e.g., Fromkin, 1971; Garrett, 1984), this speech error must be determined to be either semantic or phonological with respect to its nature. In contrast, in Dell's model, it is possible that a wrong node gets selected due to activation spreading from semantically associated nodes and phonologically associated nodes both, allowing for the intermediate nature of speech errors such as in the word exchange seen in *Let's stop* in place of *Let's start*.

The original Dell model (Dell, 1986) did not have a mechanism to explain faster reaction times for high frequency words in a lexical decision task and a naming task. To explain faster reaction times for high frequency words, frequency differences were later implemented as different activation levels of the lemma nodes (Dell, 1990). High frequency words were assumed to have higher resting activation

levels for their lemma nodes. Because of higher resting activation levels, the lemmas of high frequency words were assumed to be faster and more easily selected. Note that this way of implementing frequency differences still does not predict frequency effects on phonetic realizations. Although activation spreads two-ways in Dell's model across different linguistic levels, activation levels are used in the end only for selecting a node. In addition, the nodes in the lowest level are phonemes, not phones[3]. As a consequence, different resting activation levels are not predicted to affect phonetic realizations.

1.1.4 Levelt's model

The mechanism of spreading activation (Collins & Loftus, 1975) was integrated into a serial strictly-feedforward processing mechanism (e.g., Fromkin, 1971; Garrett, 1984) in the theory of lexical access proposed by Levelt et al. (1999) and Levelt and Wheeldon (1994) and its computational implementation WEAVER++ (Roelofs, 1997). As in Dell's model, nodes are not fully connected between different linguistic levels (e.g., syntax-morphology). However, unlike in Dell's model (e.g., Dell, 1986), activation spreads only in a forward fashion in a highly modularized hierarchical network.

Levelt's model consists of six modules: conceptual preparation, lexical selection, morphological encoding, phonological encoding, phonetic encoding, and articulation. These six modules are organized into three strata, a conceptual stratum, a lemma stratum, and a form stratum. Speech production begins with activating lexical concepts. Lexical concepts, the concepts that have corresponding words in the target language, are assumed to be non-decompositional symbols. Lexical concepts are connected with each other, and also with their corresponding lemmas. A lemma is a symbolic representation that is linked to words' inflectional features such as gender and number. Activation spreads from the lexical concepts selected

[3]Dell (1986) also proposed a phonological encoding model. In this model, the lowest nodes are phonemic features such as fricative or alveolar, and they are still abstract and distant from phonetic realizations.

to the lemma layer. The lemma node with the highest activation level is selected for further processing. At the same time, its inflectional features become available.

Once a lemma has been selected, activation flows to its morphemes. The selected morphemes in turn pass on activation to their segments and to their metrical properties. Connections between morphemes and segments are labeled for their position in the morphemes. The metrical properties only specify the number of syllables and the position of an accent. The selected segments and the metrical properties are then combined to assemble phonological words containing syllables.

Phonological words and their syllables are constructed as follows. Segments are inserted into the metrical frame from the beginning of the word. If the segment is a vowel, it will be assigned to the next nucleus position. If the segment is a consonant, the system looks ahead in the rest of the segments. If there is another vowel coming, the consonant at hand is assumed to be an onset consonant with the upcoming vowel, unless it violates the phonological rules of the language. If two consonants come in a row, they will constitute a cluster in an onset position. If the procedure makes an illegal cluster, one of the segments in the cluster will be shifted back to the preceding syllable as a coda consonant. When the system looks ahead but does not find any more vowels coming later in the rest of the strings, the consonant at hand will be assigned to a coda position in the first place. For example, for *escorting* /əskɔɹtɪŋ/, its segments and the metrical frame "$\sigma\acute{\sigma}\sigma$" should be available. First, /ə/ is assigned to the nucleus position of the first syllable, because it is a vowel. For the next segment /s/, the system looks ahead and finds another vowel (i.e., /ɔ/) coming. Therefore, /s/ is assigned to the onset position of the second syllable. The next segment /k/ is also assigned to the onset position of the second syllable, creating the consonant cluster /sk/. The consonant cluster /sk/ is legal in English, hence it is kept as it is. The next segment /ɔ/ is assigned to the nucleus position of the second syllable. The next segment /ɹ/ is a consonant. The system again looks ahead and finds /ɪ/ coming later. Therefore, /ɹ/ is first assigned to the onset of the third syllable. The next segment is /t/, also a consonant. However,

/ɹt/ is an illegal sequence in English in the onset position. As a consequence, /ɹ/ is shifted back to the coda position of the previous segment, while /t/ is assigned to the onset of the third syllable by itself. The remaining two segments /ɪ/ and /ŋ/ are assigned to the nucleus and the coda of the third syllable respectively. The end product is /ə-skɔɹ-tɪŋ/.

Following the construction of a phonological word, a gestural score is retrieved from the collection of gestural scores named "syllabary" for each syllable of the phonological word. A gestural score is still an abstract representation that only specifies speech tasks for each articulator. For example, *pan* contains [p] at its beginning. For a bilabial stop such as [p], lips must close and open. In this case, the gestural score of *pan* would specify that lips must close and open but not how. Because of coarticulation (Öhman, 1966), the articulation of [p] would be different when [p] follows [a], compared to when [p] follows [u]. Gestural scores do not specify how exactly articulators should move. Such "context-dependent" properties were assumed to be handled with and determined by an external neuromuscular execution system (Levelt et al., 1999, p. 5).

Levelt's model, however, has several problems. First, it cannot explain effects of phonological neighborhood density (Gahl et al., 2012; Vitevitch, 2002; Vitevitch & Luce, 2016; Vitevitch & Stamer, 2006). Greater phonological neighborhood density has been found to be associated with fewer speech errors (Vitevitch, 2002), shorter reaction times in the picture-naming task (Vitevitch, 2002), and shorter word duration (Gahl et al., 2012).

In addition, Levelt's model cannot deal with durational differences between homophones either. For example, Gahl (2008) investigated homophonous pairs such as *time* and *thyme* in English and found that the member of the homophonous pair with a greater frequency showed shorter duration. Such durational differences between homophones are not expected by Levelt's model, because phonetic realizations should not be affected by lexical factors such as frequency differences.

1.1.5 Connectionist models

All the classical models mentioned above are based more or less on the assumption of decompositionality. Morphologically complex words are assumed to be combinations of several morphemes such as stems and suffixes. In contrast to this assumption, Rumelhart and McClelland (1986) proposed a three-layer network to explain the past-tense formation process of English verbs without explicit rules of inflection.

This network worked exclusively with binary nodes. The input nodes of the first layer corresponded to "Wickelphones". Wickelphones are basically triples of phonemes (i.e., triphones) such as `#kA`, where `#` and `A` represent a word boundary and the diphthong eɪ respectively. Each verb was coded in terms of Wickelphones and the corresponding nodes for these Wickelphones of the verb were turned on to be 1.

The Wickelphones, however, were too specific, leading to too many units/nodes to be feasible for the computational resources in the early eighties (Rumelhart & McClelland, 1986)[4]. In this model, therefore, each phoneme was represented by four dimensions of phonetic features. The first dimension contrasted `Interrupted` (e.g., [b]), `Continuous-Consonant` (e.g., [v]), and `Vowel`. The second dimension contrasted `Stop` (e.g., [b]) vs. `Nasal` (e.g., [m]) for the first dimension being `Interrupted`, `Fricative` (e.g., [v]) vs. `Liquid/Semi-Vowel` (e.g., [l]) for the first dimension being `Continuous-Consonant`, and `High` (e.g., [ɪ]) vs. `Low` (e.g., [ɛ]) for the first dimension being `Vowel`. The third dimension contrasted `Front`, `Middle`, and `Back`, according to the articulation place of the phoneme. The fourth dimension contrasted `Voiced` vs. `Unvoiced` for consonants and `Long` vs. `Short` for vowels. Since the first and the third dimensions had three options and the second and the fourth dimensions had two options, this system represented each phoneme with $3+2+3+2=10$ units, in all 11 units, including an additional unit

[4] Assuming that English has 35 phonemes to distinguish, Rumelhart and McClelland (1986) estimated that $35^3 = 42875$ Wickelphones were necessary for the input layer only, even ignoring word boundaries.

to encode a word boundary. Using this system of encoding phonemes, a total of $11 \times 10 \times 11 = 1210$ units is sufficient to represent all possible Wickelphones.

The first layer of the model by Rumelhart and McClelland (1986) was dedicated to reduce this number of units required to encode verbs in terms of Wickelphones and was adopted to translate each Wickelphone into what was called "Wickelfeatures". Wickelfeatures are triples of phonetic features. For example, `Stop-Vowel-Nasal` would be one of the Wickelfeatures activated by the Wickelphone `kAm`. The first feature (e.g., `Stop` of `Stop-Vowel-Nasal`) comes from the first phoneme of the Wickelphone in question. The second and third features come from the second (central) and third phonemes of the Wickelphone in question respectively. With the Wickelfeatures, the number of the input nodes was cut down to 460.

The Wickelfeatures activated by the target verb through the first layer were the input to the second layer. This second layer was the main layer of the model, where learning of relations between present and past tense forms took place. The activated values of the input nodes (i.e., 1 or 0) of this second layer were multiplied with their corresponding weights to the output nodes. In other words, each output node receives the sum of the products of the input activations and the weights from the input nodes connected to the output node, its net input. The binary activation of each output node was determined probabilistically, using a logistic function.

These output nodes of the second layer were the input to the third (last) layer, which Rumelhart and McClelland (1986) called a "decoding network". The decoding network was responsible for determining that Wickelphone that explains the most of the activated Wickelfeatures. The activations of the Wickelphones were determined in such a way that the more unique Wickelfeatures a particular Wickelphone had, the more likely the Wickelphone would win the competition with other Wickelphones. However, this decoding process was not the main part of the model of Rumelhart and McClelland (1986). The performance of the model was mainly evaluated by counting the number of the Wickelfeatures that were activated

correctly in the output nodes of the second layer.

The past tense formation model by Rumelhart and McClelland (1986) sparked fierce discussions and criticism, including a very detailed criticism by Pinker and Prince (1988), pointing out several issues inherent in the model of Rumelhart and McClelland (1986). Among them was the issue of homophonous words. For example, *ring* and *wring* share the same phonemes, namely /rɪŋ/ (Pinker & Prince, 1988). Nevertheless, *ring* is irregular, while *wring* is regular[5]. The model of Rumelhart and McClelland (1986) simply predicts a corresponding past-tense form from the stem form of a certain verb in terms of their phonetic features. As a consequence, homophonous verbs necessarily are associated with the same output, regardless of the (ir)regularity of their past tense forms. This, of course, is an inevitable problem for any model deriving a past tense form from a present-tense form without having access to semantics.

Another problem that was pointed out concerned verbs for which the present and past tense forms are identical — an identity mapping (Pinker & Prince, 1988). Pinker and Prince (1988) claimed that the transformation between identical forms (i.e., no change such as `#me` → `#me`) should be easier than non-identical mappings (e.g., `#me` → `xyz`). Nevertheless, in the model of Rumelhart and McClelland (1986), no special status was given to the identity mapping, and, as a consequence, the model would predict that the mapping between the same form and the mapping from one form to another completely different form would be the same in difficulty. Although Rumelhart and McClelland (1986) observed that the no-change verbs, where the stem and the past tense forms are identical such as *cut*, were produced more accurately compared to other irregular verbs, Pinker and Prince (1988) claimed that this good performance on the no-change verbs was merely an artifact of the use of the Wickelfeatures as the validation method.

In response to these criticisms, MacWhinney and Leinbach (1991) improved the model of Rumelhart and McClelland (1986) mainly by increasing the number

[5] Although Pinker and Prince (1988) explained that the past tense *wring* is *wringed*, there is in fact variability in its past tense forms, namely *wringed*, *wrang*, or *wrung*.

of layers (by adding hidden layers), in addition to improving the input and output representations. Different definitions of the input and the output were implemented in response to one of the claims by Pinker and Prince (1988) that Wickelphones were problematic. However, Pinker and Prince (1988) already expected this line of argumentation, claiming that although increasing the number of layers would probably lead to better performance of the model, it would obscure the distinction between the Parallel Distributed Processing (PDP) models and the classical rule-based theories. Pinker and Prince (1988) pointed out that several properties of the model by Rumelhart and McClelland (1986) were actually motivated by the rule-based theories and therefore the PDP models including the model of Rumelhart and McClelland (1986) were merely implementations of the rule-based theories. For example, the model of Rumelhart and McClelland (1986) mapped the stem form onto the past-tense form. The use of the concepts of the stem and the past-tense inflected form was motivated by the property of the rule-based account, in which the stem was combined with a suffix to produce an inflected form. An increase in the number of layers would help the model capture certain patterns equivalent to the "rules" such as "change the stem vowel from /i/ to /ɪ/ and add /t/ at the end except when the word final consonant is already /t/" (e.g., *meet* $\rightarrow$ *met*). This would undermine the claim by Rumelhart and McClelland (1986) that the past tense formation could be explained without rules, according to Pinker and Prince (1988). However, with current advances in deep learning, it has become clear that multi-layer networks are very good at capturing many kinds of regularities and sub-regularities, while at the same time profiting from many low-level statistical correlations that are beyond the scope of the kind of rules envisioned by Pinker and Prince.

Several issues pointed out by Pinker and Prince (1988) such as mentioned above can actually be readily resolved, once semantics is properly taken into consideration. Pinker and Prince (1988) argued that semantics should not play a role, on the basis of arguments such as that if semantics were to play a role, verbs with

similar meanings should have similar past tense forms (e.g., *hit*/*hit*, *strike*/*struck*, vs. *slap*/*slapped*). Since then, it has become clear that semantics does play a much more important role (see, e.g., Baayen & Moscoso del Prado Martín, 2005; Heitmeier & Baayen, 2020; Ramscar, 2002).

1.1.6 Discriminative learning models

The connectionist models such as the one by Rumelhart and McClelland (1986) focused on the mapping from one from to another, following common practise in linguistics, a practise that in turn was inspired by pedagogical grammars (Blevins, 2016). In contrast, the discriminative lexicon model (DLM) developed by Baayen et al. (2019) focuses on the relation between form and meaning. This model, which addresses both comprehension and production, builds on an earlier model that was developed only for comprehension, the naive discriminative learning model (NDL: Baayen et al., 2011).

The NDL model made use of a simple learning rule, proposed by Rescorla and Wagner (1972), known as the Rescorla-Wagner learning rule (see also Rescorla, 1988). This learning rule incrementally estimates association strengths (i.e., weights), based on co-occurrences of cues and outcomes. Cues are word form features and outcomes are one-hot encoded word meanings. For example, cues can be letter bigrams or phone trigrams. Furthermore, the meanings of complex words can be represented by multiple semantic nodes, including not only nodes for lexical meanings but also nodes for grammatical meanings.

The Rescorla-Wagner rule is used to learn the weights between form cues and semantic outcomes. When a particular cue is absent, weights on the connections from this cue to the outcomes are not updated. When a particular cue co-occurs with a particular outcome, the association strength between the cue and the outcome is strengthened. When a particular cue is present but a particular outcome is absent, the cue is not so likely to be a good cue for the outcome, and accordingly the association strength from this cue to the outcome is weakened. These updates

of association strengths are modified by presence of other competing cues. When there are many other competing cues in addition to the target cue of interest, the positive update of the association strength due to the co-occurrence of the cue with a certain outcome will be smaller, while the association strength will be updated negatively more strongly, if a certain outcome is absent. These weight-updating rules are summarized as the following equation, which specifies the change in connection strength:

$$\Delta V_i^t = \begin{cases} 0 & \text{if } \textsc{absent}(C_i,t) \\ \alpha_i\beta_1(\lambda - \sum_{\textsc{present}(C_j,t)} V_j) & \text{if } \textsc{present}(C_j,t) \;\&\; \textsc{present}(O,t) \\ \alpha_i\beta_1(0 - \sum_{\textsc{present}(C_j,t)} V_j) & \text{if } \textsc{present}(C_j,t) \;\&\; \textsc{absent}(O,t) \end{cases}$$

where ΔV_i^t denotes the amount of the update applied to the association strength V of a certain cue C_i to the outcome O at the time point t, and $\textsc{present}(X,t)$ and $\textsc{absent}(X,t)$ denote the presence and the absence of the cue or outcome at the time t respectively. $\sum_{\textsc{present}(C_j,t)} V_j$ represents the sum of the associations of all the cues present at the time point t to the outcome O.

This NDL model with the Rescorla-Wagner learning rule (Rescorla, 1988; Rescorla & Wagner, 1972) is similar to the connectionist model of Rumelhart and McClelland (1986) in the sense that both make use of a one-layer network[6]. Furthermore, both models do not implement discrete rules that operate on stems and exponents. The critical difference is that NDL maps forms onto meanings, while the connectionist models map forms (e.g., the stem form) onto other forms (e.g., the past tense form). A number of studies using NDL has found that the degree to which word meanings are supported by their forms is predictive for processing measures such as lexical decision latencies (Baayen et al., 2011; Baayen &

[6]The model by Rumelhart and McClelland (1986) technically has three layers. However, the first layer is dedicated to convert Wickelphones to Wickelfeatures, and the third later is a fixed network to decode (predicted) Wickelfeatures back to Wickelphones. Therefore, the main "learning" occurs only in the middle one layer.

Smolka, 2020), self-paced reading times (Baayen et al., 2011), and acoustic durations (Tomaschek, Plag, et al., 2019; Tucker et al., 2019).

A downside of NDL is that it treats meanings as symbols, implemented as orthogonal one-hot encoded vectors. However, *child* should be closer in semantics to *kid* than *universe*, since *child* and *kid* both refer to a young person while *universe* does not even refer to animals or creatures. Since different meanings are orthogonal to each other in NDL, a semantic distance (e.g., Euclidean distance) between *child* and *kid* is the same as that between *child* and *universe*.

This issue was resolved by adopting real-valued semantic vectors as approximations of the meanings of words, resulting in the discriminative lexicon model (DLM, Baayen et al., 2019). The vectors representing word meanings can be simulated, or instead empirical corpus-based semantics can be used, generated with methods such as word2vec (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013). When empirical vectors are used, lexical similarities such as those of *child* and *kid* are properly represented in the model. The core engine of the DLM is the way in which mappings between cues and outcomes are obtained, using the mathematics of multivariate multiple regression. In other words, the DLM takes semantic representations as a given, and also takes form representations as a given, and then uses learning to obtain mappings from form to meaning, and from meaning to form. For comprehension, the equation to be solved is

$$\mathbf{CF} = \mathbf{S}. \tag{1.1}$$

In (1.1), $\mathbf{C}$ is a word $\times$ cues matrix specifying words' forms. In other words, rows of $\mathbf{C}$ define words' forms, and columns of $\mathbf{C}$ specify letter or phone n-grams. For example, using triphones, the word forms of *hand* and *band* can be represented with a matrix $\mathbf{C}$ as follows:

$$\mathbf{C} = \begin{array}{c} \\ \text{hand} \\ \text{band} \end{array} \begin{array}{c} \begin{array}{cccccc} \text{\#ha} & \text{han} & \text{\#ba} & \text{ban} & \text{and} & \text{nd\#} \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix} \end{array} \tag{1.2}$$

Instead of using triphones to represent words, real-valued vectors can be used, derived, for instance, from a word's spectrogram (Shafaei-Bajestan et al., 2021).

S in the equation 1.1 above represents the matrix specifying words' semantics. Its rows pertains to words and its columns specify semantic dimensions. The row vectors of **S** are words' semantic vectors of words (also known as word embeddings), which can be constructed by any method of creating word embeddings such as word2vec (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013), fastText (Bojanowski et al., 2017), simulated semantic vectors (Baayen et al., 2018), or the NDL-based estimation (Baayen et al., 2019). **S** contains real-values and may look like the following (in which the numbers are at random just for purpose of illustration):

$$\mathbf{S} = \begin{array}{c} \\ \text{hand} \\ \text{band} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{S1} & \text{S2} & \text{S3} & \text{S4} & \cdots \end{array} \\ \begin{bmatrix} 0.43 & 0.29 & -0.88 & -0.03 & \cdots \\ -0.35 & 0.22 & -0.49 & 0.17 & \cdots \end{bmatrix} \end{array} \tag{1.3}$$

F represents association strengths between each form dimension and each semantic dimension. This matrix can be estimated in the following way, using the normal equations of regression:

$$\mathbf{F} = (\mathbf{C}^\top \mathbf{C})^{-1} \mathbf{C}^\top \mathbf{S} \tag{1.4}$$

The resulting mapping matrix **F** represents the *endstate of learning* (Chuang & Baayen, 2021; Heitmeier et al., 2021; Shafaei-Bajestan et al., 2021). Therefore, **F** can conceptually be viewed as the knowledge about associations between forms and meanings accumulated thanks to infinite learning experience with the data. The

weight matrix **F** can also be learned (estimated) incrementally, using the learning rule of Widrow and Hoff (1960):

$$\mathbf{F}^t = \mathbf{F}^{t-1} + \mathbf{c}^\top(\mathbf{s} - \hat{\mathbf{s}})\eta, \tag{1.5}$$

where $\mathbf{F}^t$ represents the updated association strengths between forms and meanings, and $\mathbf{F}^{t-1}$ pertains to the association strengths before the update. The vector **c** is a form vector, where present cues are marked as 1 otherwise 0. The vectors **s** and $\hat{\mathbf{s}}$ are the "correct" (or gold-standard) and the predicted semantic vectors respectively. η is a parameter governing the learning rate. Simply put, the equation 1.5 states that the associations strengths between forms and meanings are updated according to the errors between the correct and predicted semantic vectors. If the predicted semantic vector has too small a value for a certain semantic dimension, the pertinent association strength gets updated upwards. In contrast, if a certain association strength was too high than it should be, then this association strength will be updated downwards.

One important difference between the two ways of estimating the mapping **F** from the psycholinguistic perspective is that the endstate of learning does not reflect differences in frequency of occurrence. Since the endstate of learning estimates the equilibrium of association strengths, high frequency and low frequency words are treated in the same way (Heitmeier et al., 2021). The insensitivity to frequency differences is avoided by using frequency-informed endstate learning (for more detail, see Heitmeier et al., 2022).

F constitutes the "learning" part of a 'linear discriminative learning' (LDL) mapping. Given **F**, we can predict the semantic vector corresponding to a form vector as follows:

$$\mathbf{c}_i\mathbf{F} = \hat{\mathbf{s}}_i, \tag{1.6}$$

where $\mathbf{c}_i$ is the form vector of the i-th word, and $\hat{\mathbf{s}}_i$ is the predicted semantic vector

of the *i*-th word. The predicted meanings of all words jointly can be expressed in terms of matrix multiplication as below:

$$\mathbf{CF} = \hat{\mathbf{S}}, \tag{1.7}$$

where each row of $\mathbf{C}$ and $\hat{\mathbf{S}}$ corresponds to the comprehension of each word.

So far, cues were forms, and outcomes were meanings, as in NDL. However, the DLM also sets up mappings from meanings to forms, thereby implementing a first step in the process of speech production. The mapping (weight) matrix for production can be estimated as follows:

$$\mathbf{G} = (\mathbf{S}^\top \mathbf{S})^{-1} \mathbf{S}^\top \mathbf{C}, \tag{1.8}$$

Using this weight matrix, the predicted form vectors corresponding to words' semantic vectors are obtained straightforwardly:

$$\mathbf{SG} = \hat{\mathbf{C}} \tag{1.9}$$

The predicted form vectors specify the amount of support that individual n-phones receive from words' meanings. In order to properly order the triphones into the sequence required for articulation, a subsequent algorithm is used. As this algorithm is not relevant to the present thesis, the reader is referred to Baayen et al. (2019) and Luo et al. (2021) for further information.

As is the case for NDL, the representations for words' forms in the DLM are not informed by any linguistic units such as stems, exponents, or words. The only units used to define forms in production are n-phones (e.g., triphones). Nevertheless, LDL has been validated language-internally by predicting forms including unseen forms (Baayen et al., 2018; Baayen et al., 2019; Shafaei-Bajestan & Baayen, 2018; Shafaei-Bajestan et al., 2021) and also externally in terms of predictivity for processing measures such as lexical decision latencies and acoustic durations (Baayen et al., 2019; Chuang et al., 2019). For example, in Baayen et

al. (2019), English content words are correctly produced by the model from letter trigrams with an accuracy of more than 99%. Inflected words were also correctly predicted with an accuracy of 92%. Moreover, using 10-fold cross validation, the model predicted 62% of unseen inflected forms correctly. Similarly, derived words are predicted correctly with an accuracy of 99% when all the derived words were included in the training. Even when a part of all the derived words were withdrawn from training (i.e., 10-fold cross validation), an accuracy of 75% was retained.

The DLM model also provides a framework for understanding speech errors. Unlike the past-tense model of Rumelhart and McClelland (1986), the DLM model produces errors that resemble the kind of errors that actual speakers might make (see, e.g. Chuang et al., 2020). Speech errors such as *slicely thinned* (instead of *thinly sliced*, which provide strong prima facie evidence for morphemes being exchanged, are straightforward to generate in the DLM. DLM locates the source for this kind of error at the semantic level: instead of combining the semantic vectors of *thin* and *-ly*, and those of *slice* and *-ed*, the vectors of *slice* and *-ly* are combined, and likewise those of *thin* and *-ed*. The resulting semantic vectors then straightforwardly produce the forms *slicely* and *thinned*.

In addition, the DLM model does not suffer from the criticism by Pinker, which was mentioned above. One of the criticism involved homophonous English verbs, one of which is regular and the other is irregular in terms of their past tense forms (e.g., *ring - rang* vs. *wring - wringed*). The DLM model predicts that any form, including regular/irregular past tense forms, is motivated by words' meanings. Homophonous verbs differ in semantics. Therefore, it is straightforwardly predicted without any problem that one of a homophonous pair of present-tense forms can have a regular past tense form while the other has an irregular past tense form. Importantly, the DLM does not derive a past-tense form from a present-tense form, forms are derived from their meanings. Another point of the criticism involved the identity mapping, where a present tense verb form and its corresponding past tense form are identical (e.g., *cut*). In the DLM model, past-tense forms are predicted

primarily by semantics, as is the case for any word and any form, not by their corresponding present-tense forms. As a consequence, the DLM model predicts that the identity mapping should basically be the same as, or at least very similar to, other not-identity mappings, as was the case for the model of Rumelhart and McClelland (1986).

1.2 Effects of morphological boundaries

Serial feed-forward modular models such as the one by Levelt et al. (1999) integrate morphemes (or rather, exponents) into the theory. Lemmas selected after the completion of conceptualization activate morphemes in turn. The selected morphemes provide pointers to their constituent phonemes. These phonemes are subsequently combined into syllables, forming phonological words (Levelt et al., 1999). In this theory, the presence or absence of a morpheme boundary is not preserved in the form representations. As a consequence, words that share the same phonemes and syllables are predicted to be pronounced in the same way, even though one word may have an internal morpheme boundary whereas the other doesn't.

In contrast to this prediction, a number of studies have reported effects of a morpheme boundary on phonetic realizations (Hay, 2007; V. G. Li et al., 2020; Plag & Ben Hedia, 2018; Seyfarth et al., 2017; Smith et al., 2012; Song et al., 2013; Strycharczuk & Scobbie, 2016; Sugahara & Turk, 2009). For example, Seyfarth et al. (2017) looked into homophonous pairs (e.g., *lap+s* vs. *lapse*) and found that stems were significantly longer in the pre-morpheme-boundary condition. In other words, pre-morpheme-boundary segments were shown to be phonetically enhanced.

A similar effect was also observed for prefixes (Hay, 2007; Plag & Ben Hedia, 2018; Smith et al., 2012). Hay (2007) approached effects of a morpheme boundary from the perspective of morphological boundary strength. Using relative frequency (word frequency divided by lemma frequency), Hay (2007) argued that lower relative frequency facilitates morphological parsing. Furthermore, she argued that low probability phone transitions constitute strong morphological boundaries, resulting in more enhanced phonetic realization.

Whereas Hay (2007) investigated total duration of the English prefix *un-*, Smith et al. (2012) studied the duration of individual segments in prefixes and observed that it was mainly the vowel in a prefix (e.g., *dis+tasteful*) that was enhanced phonetically before the morpheme boundary. Also in the domain of articulatory re-

alizations, a similar enhancement effect has been observed (V. G. Li et al., 2020; Song et al., 2013). For example, longer intervals between maximal constrictions of the tongue were found when a morpheme boundary is involved (V. G. Li et al., 2020).

A recent study (Baayen et al., 2019) suggested that effects of the morphological boundary may be based on form-meaning relationships of morphologically complex words. Baayen et al. (2019) created word-embeddings (semantic vectors) using Naive Discriminative Learning (NDL: Baayen et al., 2011) and trained a Linear Discriminative Learning model (LDL: Baayen et al., 2019) to learn associations between sublexical forms (i.e., trigrams) and meanings of the words. They found that the branching segment, namely the final segment of the stem (e.g., *d* in *blend*, *blends*, *blended*, and *blending*), was longer when the transition from the branching segment to the next segment was not strongly supported by the word's semantics. This finding suggests that effects of the morphological boundary can be semantic in nature.

1.3 Effects of frequency

Frequency effects in the context of psycholinguistics were first discovered by (Oldfield & Wingfield, 1965). Their study found that naming latency is slower for lower frequency words. Effects of frequency on reaction times in a lexical decision task and a naming task were not within the scope of the models by Fromkin (1971) and Garrett (1984), which mainly aimed at explaining possible speech errors. Dell's model later incorporated different resting activation levels for lemma nodes (Dell, 1990) to explain shorter response times for high frequency words. Different activation thresholds were also used by Jescheniak and Levelt (1994) and also by Levelt's model (Levelt et al., 1999)[7].

[7]The computational implementation of Levelt's model, WEAVER++, implemented frequency effects by means of verification times. In this model, each selection of a lexical item has to be verified by being compared to the node in the immediately upper level. Higher frequency words are assumed to be verified faster and therefore afford shorter response times.

After the Oldfield et al. study, effects of frequency have been documented for many measures of lexical processing, such as lexical decision and naming latencies (Baayen et al., 1997; Baayen et al., 2006; Baayen et al., 2002; Bertram et al., 2000; Forster & Chambers, 1973; Gardner et al., 1987; Rubenstein et al., 1970; Scarborough et al., 1977; Schreuder & Baayen, 1997; Whaley, 1978; Wurm et al., 2006), speech errors (Gordon, 2002; Harley & Bown, 1998; Vitevitch, 1997), acoustic characteristics (Aylett & Turk, 2004; A. Bell et al., 2009; A. Bell et al., 2002; Dinkin, 2008; Jurafsky et al., 2001; Munson & Solomon, 2004; Pluymaekers et al., 2005b), and tongue movements (Lin et al., 2011; Tomaschek, Arnold, et al., 2018; Tomaschek, Tucker, et al., 2018; Tomaschek et al., 2013).

The classical speech production models introduced above, i.e. Fromkin's model, Garrett's model, Dell's model, and Levelt's model, cannot explain frequency effects on phonetic realizations, although at least Dell's model and Level's model have some mechanism to explain frequency effects on reaction times. In Dell's model, higher resting activation levels for high frequency words may have some influence on the phoneme nodes further down in the processing mechanism through the morpheme nodes. However, phoneme nodes represent phonemes, not phones. Selected phonemes must be translated into articulatory gestures later. Furthermore, phonemes are either selected or not selected. The continuous nature of frequency effects is not well served by such a discrete selection mechanism. Similarly, Levelt's model limits word frequency effects to the word-form selection stage, which excludes the possibility that different frequencies of occurrences of words might affect the phonetic realizations of these words. It might be argued that it is syllable frequency that co-determines phonetic realizations. However, in Levelt's model, gestural scores (accessed through the syllabary) are assumed to be abstract and context-free, only specifying speech tasks (e.g., "close" for lips), while leaving out how a certain speech task is actually carried out by an external system (called "a neuromuscular execution system") (Levelt et al., 1999, p. 31). As a consequence, systematically different phonetic realizations cannot be predicted

by these classical models.

Furthermore, several different kinds of frequency have been investigated, including whole-word (surface) frequency (Aylett & Turk, 2004; Baayen et al., 1997; Baayen et al., 2006; A. Bell et al., 2009; A. Bell et al., 2002; Bertram et al., 2000; Dinkin, 2008; Jurafsky et al., 2001; Lin et al., 2011; Munson & Solomon, 2004; Pluymaekers et al., 2005b; Wurm et al., 2006), lemma frequency (Baayen et al., 1997; Baayen et al., 2011; Bertram et al., 2000; Gahl, 2008; Lohmann, 2018a, 2018b), relative frequency of whole-word and stem (Hay, 2007; Hay, 2003; Stein & Plag, 2022), segment frequency (Van Son & Van Santen, 2005), phrase (multiple-word) frequency (Arnon & Cohen Priva, 2013), and constituent frequency (Duñabeitia et al., 2007; Schmidtke et al., 2021).

Frequency counts have to be interpreted against the background of the size of the corpus from which counts are calculated. Frequency divided by the corpus size is known as a word's prior probability (Jurafsky et al., 2001) or as its relative frequency (Jurafsky et al., 2001; Tomaschek, Tucker, et al., 2018). Furthermore, several refinements of frequency-based estimates of probability have been proposed, such as conditional probability (predictability) (Aylett & Turk, 2004, 2006; Jurafsky et al., 2001), paradigmatic probability (M. J. Bell et al., 2021; Cohen, 2014; Kuperman et al., 2007; Tomaschek et al., 2021), amount of information (surprisal) (Brandt et al., 2021; Cohen Priva, 2015; Kuperman et al., 2007; Malisz et al., 2018; Van Son & Van Santen, 2005), and entropy (Baayen et al., 2006; Kuperman et al., 2007; Moscoso Del Prado Martín et al., 2004).

With respect to acoustic realization, words with a higher frequency of use are typically realized with shorter durations A. Bell et al. (2002) and more centralized formant structures (Aylett & Turk, 2006). Gahl (2008) and Lohmann (2018b) showed for homophone pairs such as *time* and *thyme* that the less frequent homophone is realized with longer spoken word duration. Pluymaekers et al. (2005b) focused on the durations of prefixes and suffixes and reported shorter durations for some of the affixes when they occurred in higher frequency words. Dinkin (2008)

investigated short vowels in English (e.g., ɪ, ɛ, æ, ʌ, ʊ) and demonstrated that in higher frequency words front vowels show lower F2 and back vowels show higher F2, indicating that in higher frequency words these vowels are more centralized. To probe articulation itself, Lin et al. (2011) used ultrasound imaging to compare the height of the tongue tip during the pronunciation of /l/ in a /(C)(C)VlC/ context (e.g., *milk*) for carrier words of different frequencies. They observed that the tongue tip was located lower in the oral cavity for higher frequency words, again indicating articulatory reduction.

The finding that higher-frequency words tend to undergo more articulatory reduction, has been explained in terms of syntagmatic redundancy or predictability. Jurafsky et al. (2001) introduced the conditional probability of the current word given the previous word as a predictor for spoken word duration and found that higher values of the conditional probability measure were associated with shorter spoken word duration. Aylett and Turk (2004) defined syllable-level trigram probability as a further measure gauging syntagmatic redundancy. Syllable-level trigram probability is the probability of the target syllable given the preceding two syllables (regardless of word boundaries). Aylett and Turk (2004) reported more phonetic reduction for words with higher syllable-level trigram probability. These studies all suggest that the words that are syntagmatically more predictable and redundant are reduced phonetically. These reduction phenomena all are well described by the smooth signal redundancy hypothesis of Aylett and Turk (2004), according to which high probability highly redundant words are pronounced with shorter durations and more centralization in order to obtain a more smoothly evolving speech signal.

The reduction effect of syntagmatic predictability has also been found in the form of syntactic structures/contexts. Gahl and Garnsey (2004) investigated whether match/mismatch of the verb-bias and the syntactic structure the verb occurs in affected pronunciations of the verb. In English, some verbs are biased to the direct-object structure, e.g., *confirm the date*, while others are biased to the

sentential complement structure, e.g., *suggest the date should be changed*. Gahl and Garnsey (2004) found that verb-final stops were more likely to be deleted when the verb and the syntactic structure matched than when they mismatched. Direct-object verbs were significantly longer when they were used with the sentential complement, compared to when they were followed by direct objects as expected by the verb. In addition, post-verbial silences were longer in bias-violating than bias-matching sentences.

In contrast to the greater probability of phonetic reduction that has been widely reported to go hand in hand with greater frequency of use, an opposite frequency effect tied to phonetic enhancement has also been reported. Kuperman et al. (2007) looked into the interfixes occurring in Dutch noun-noun compounds. Krott et al. (2001) had shown that the choice of an interfix in a compound is based on the distribution of interfixes that follow the left constituent. The constituent with the highest probability in this mini-paradigm defined by the left constituent is the most likely to appear in novel compounds, and is the easiest one to process (Krott et al., 2007). Based on this research, Kuperman et al. (2007) reasoned that the the most probable interfix should have the shortest duration. However, the opposite was observed: the more probable an interfix is in its mini-paradigm, the longer its acoustic duration is. This led Kuperman et al. (2007) to propose the paradigmatic signal enhancement hypothesis according to which greater paradigmatic support leads to phonetic enhancement rather than phonetic reduction.

Since the finding by Kuperman et al. (2007), several studies have replicated the same effect of frequency and paradigmatic probability. For example, M. J. Bell et al. (2021) investigated consonant duration at compound-internal morphological boundaries and found that greater probability of consonants at the boundary following the first noun go hand in hand with longer consonant duration. In addition, Cohen (2014) investigated the duration of the English verbal plural suffix (i.e., *-s*) and found that the suffix duration was longer when the present third-person singular form was more likely than the corresponding plural (stem) form, which was

gauged by relative frequency of the verb's singular frequency divided by the plural frequency. Tucker et al. (2019) investigated English irregular past-tense verbs and defined paradigmatic probability as the number of irregular verbs sharing the same vowel alternation pattern between the present and past tense forms. In addition, Tucker et al. (2019) also defined its learning-based alternative, called the vowel-tense activation, derived from Naive Discriminative Learning (NDL: Baayen et al., 2011). The vowel-tense activation gauged how much the tense was supported by the diphones containing the stem vowel of the verb. Both measures were assumed to capture the amount of support for a certain vowel alternation and therefore paradigmatic (un)certainty. These two measures both showed U-shaped effects with the stem vowel duration being longer for their larger values (Tucker et al., 2019). The authors of the study interpreted these results as a partial support for the paradigmatic signal enhancement hypothesis. Also in the articulatory domain, Tomaschek et al. (2021) found lower tongue trajectories for the [ɑ] vowel in the stem of English verbs when these verbs had a higher frequency of occurrence. Since [ɑ] is an open low vowel, the finding of a lower tongue trajectory for higher frequency words indicates articulatory enhancement.

These enhancement effects of frequency have been explained as a consequence of resolving uncertainty in morphological paradigms. A more probable paradigmatic alternative reduces this uncertainty and thereby affords phonetic strengthening.

Chapter 2

Articulatory effects of frequency modulated by semantics

This chapter will be published as: Motoki Saito, Fabian Tomaschek, R. Harald Baayen. Articulatory effects of frequency modulated by semantics. In Marcel Schlechtweg (ed.), Interfaces of Phonetics (Phonology and phonetics series). De Gruyter Mouton.

Abstract: This chapter provides an overview of three studies addressing the role of frequency in speech production. While frequency has often been observed to be correlated with phonetic reduction, as evidenced by shorter durations and more vowel centralization for higher-frequency words, some studies have reported phenomena for which a higher frequency appears to give rise to phonetic enhancement. These opposite effects of frequency have thus far resisted a consistent, theoretically well-motivated, explanation. The first case study replicates the effect of phonetic enhancement, previously observed with EMA, using ultrasound recordings. The second case study looks into the possibility that morphological complexity codetermines phonetic enhancement, using EMA. The third case study provides evidence that words' meanings, gauged with distributional semantics, play an important role in shaping phonetic enhancement and reduction.

2.1 Introduction

Frequency is one of the most extensively investigated variables used for probing lexical processing (see, e.g., Baayen et al., 2016, for an overview). More frequent units (e.g., words) have repeatedly been found to have shorter acoustic duration (Aylett & Turk, 2004; A. Bell et al., 2009; A. Bell et al., 2002; Gahl, 2008; Malisz et al., 2018; Pluymaekers et al., 2005a, 2005b) and to have more centralized formant realizations (Aylett & Turk, 2006; Dinkin, 2008; Wright, 2004). The phonetic reduction associated with higher frequency has been explained as a consequence of syntagmatic predictability (Aylett & Turk, 2004): higher-frequency words are less informative and therefore realized with more reduced forms.

In contrast, greater frequency of occurrence has also been reported to be associated with phonetic enhancement (M. J. Bell et al., 2021; Cohen, 2014; Kuperman et al., 2007; Tomaschek, Tucker, et al., 2018; Tomaschek et al., 2021). These studies observed that segments tend to be longer in duration and to be articulated with more peripheral tongue positions when a word's frequency or probability in its paradigm is higher, yielding more discriminative phonetic characteristics. A paradigm is a set of words that are morphologically related to each other. Kuperman et al. (2007) focused on paradigmatic probability and observed that greater paradigmatic probability leads to phonetic enhancement. Furthermore, Tomaschek, Tucker, et al. (2018) reported evidence that higher frequency words were realized with clearer articulations. Tomaschek, Tucker, et al. (2018) argue that these clearer articulations are the result of enhanced motor control for higher frequency words: opportunities for learning the planning and execution of the articulatory gestures involved in producing a word present themselves more often for higher frequency words as compared to lower frequency words. According to Tomaschek et al. (2021), similar effects of enhancement are also present for inflected words that have a higher probability in their inflectional paradigm.

A recent study by Gahl and Baayen (2022) proposed that the reduction and enhancement effects of frequency may indeed be orthogonal and capture two dif-

ferent aspects of the speech production process. According to Gahl and Baayen (2022), phonetic reduction as a function of frequency arises at the level of the utterance, where words enter into syntagmatic relations, and where a word's probability hinges on the preceding words in the utterance and in the preceding discourse. Since higher frequency words tend to be more predictable, and hence are less informative, the greater phonetic reduction observed for higher frequency words is well explained by the smooth signal hypothesis proposed by Aylett and Turk (2004) and similar explanations in subsequent studies (A. Bell et al., 2009; A. Bell et al., 2002; Gahl, 2008; Pluymaekers et al., 2005a, 2005b).

Words do not only play a role as units of meaning in utterances and discourse but also as the units that have to be articulated. According to Gahl and Baayen (2022), it is during the mapping of a word's meaning onto its form that frequency, as a measure of experience and articulatory practice, gives rise to articulatory enhancement. As a consequence of learning associations between meanings and forms, certain meanings are more strongly associated with certain forms than others. According to Gahl and Baayen (2022), when a word's semantics provides stronger support for that word's form, there is more evidence for that form, which leads to enhanced articulation. Conversely, when there is no support from the semantics for a word form, in the limit, it is not articulated at all and has zero duration. This concept is related to the paradigmatic signal enhancement hypothesis (e.g., Kuperman et al., 2007). When a choice has to be made as to which of a set of allomorphs has to be selected, the one with the greater probability will be realized with more enhancement.

Given that the reduction and enhancement effects of frequency are potentially orthogonal, it remains unclear why reduction is found for some cases and enhancement for other cases. If the proposal by Gahl and Baayen (2022) is correct and these two directions of frequency effects are orthogonal, there are always two forces at work, and how exactly they work out jointly is a matter for further empirical investigation.

What then modulates the balance of the two effects? One possible factor is morphological complexity. When enhancement effects of frequency (or frequency-based measures such as probability) are observed, the items under investigation tend to be morphologically complex words. For example, enhancement effects of frequency have been demonstrated for interfixes of compounds (Kuperman et al., 2007), inflectional suffixes (Cohen, 2014), geminates across a morpheme boundary (M. J. Bell et al., 2021), and stem vowels of inflected verbs (Tomaschek, Tucker, et al., 2018; Tomaschek et al., 2021).

In contrast, reduction effects of frequency have been observed frequently in studies focusing on monomorphemic words (Lin et al., 2011; Wright, 2004) or morphologically simple and complex words mixed (Aylett & Turk, 2004; A. Bell et al., 2009; A. Bell et al., 2002; Dinkin, 2008; Gahl, 2008). However, there are also studies investigating morphologically complex words that reported enhancement effects or mixed results (Pluymaekers et al., 2005b). Thus, it remains by and large unclear how morphological status (i.e., morphologically simple vs. complex) influences frequency effects. This chapter presents three studies that address the question of whether, and if so, how, frequency effects are modulated by morphological complexity. The first study focuses on replicating the enhancement effect of frequency using ultrasound, the second study investigates the interaction of frequency morphological status using electromagnetic articulography (EMA), and the third study brings in word meaning, gauged with distributional semantics, as a new factor co-determining articulatory enhancement and reduction.

The first study follows up on earlier work using articulography with the aim of clarifying the role of frequency for the production of morphologically simple and complex words. The second study we report here made use of a corpus of German spontaneous speech using electromagnetic articulography (Arnold & Tomaschek, 2016). This second study revealed that segments preceding a morphological boundary were associated with more enhanced articulatory realizations. According to classical models of speech production (e.g., Levelt et al., 1999), nei-

ther whole-word frequency nor morphological status are expected to affect articulation, as prior to articulation, phones in a word are assembled from morphemes and bundled into syllables. The third study we review in this chapter argues that the effects of frequency and morphological status emerge naturally once the relation between meaning and form is taken into account. The computational framework that we use to predict the effect of meaning on form is that of the discriminative lexicon (Baayen et al., 2019). In the general discussion, we present a proposal for how the effect of frequency-as-information and the effect of frequency-as-articulatory-practice can be integrated within this framework.

2.2 Enhanced articulations of the stem vowel by frequency modulated by inflectional suffixes

In this section, we follow up on the study of Tomaschek, Tucker, et al. (2018). Their study had German speakers pronounce German inflected verbs with suffixes sharing the same place of articulation, namely [t] vs. [(ə)n], together with their corresponding pronouns (e.g., *sie malt* 'she paints' vs. *sie malen* 'they paint'). The stem vowel was kept the same (i.e., [aː]). Tongue positions were recorded by EMA. They found that the stem vowel [a(ː)] was articulated with the strongest tongue tip/body lowering in the middle of the vowel for high and low frequency words. Because they focused on the stem vowel [aː], lowered tongue trajectories indicate articulatory enhancement. For medium frequency words, the least lowering (i.e., the highest/smoothest tongue trajectories) was observed.

In addition, Tomaschek, Tucker, et al. (2018) found an earlier initiation of coarticulatory raising of the tongue tip/body, anticipating the upcoming suffix (e.g., [t] or [n]), for high frequency words, compared to low frequency words. Based on these observations, Tomaschek, Tucker, et al. (2018) suggested that high frequency words were articulated in such a way that the tongue tip/body reaches the lowest point of the articulation and nevertheless rises back to high tongue positions very

quickly. They interpreted the findings as a result of kinematic improvement, the idea being that also for articulation, "practice makes perfect". These effects of frequency were limited to the suffix condition [t], even though the suffixes shared the same place of articulation.

Following up on Tomaschek, Tucker, et al. (2018), we carried out a similar experiment to that by Tomaschek, Tucker, et al. (2018) to look into whether, when ultrasound is used, this effect of frequency also emerges and whether the effect of frequency is systematically modulated between the suffix conditions. To this end, we selected 153 German verbs (word types) with the same criterion as adopted in Tomaschek, Tucker, et al. (2018). 122 of these verbs were inflected with the suffixes [-t] and [-n]. The other 31 of them were combined with the suffix [-n]. They were monosyllabic when combined with the suffix [-t], e.g., *sie malt* [ziː maːlt], and disyllabic when combined with the suffix [-(ə)n], e.g., *sie malen* [ziː maːl(ə)n]. The inflected forms of the target verbs were combined with the pronoun *sie* [ziː]. The pronoun *sie* can be the third-person singular pronoun, e.g., *sie malt* [ziː maːlt] 'she paints', and also the third-person plural pronoun, e.g., *sie malen* [ziː maːl(ə)n] 'they paint'.

For each of these target inflected verbs, frequency was obtained from the SdeWac corpus of German (Faaß & Eckart, 2013). The range of frequency in the present study is roughly compatible with that of Tomaschek, Tucker, et al. (2018). There was no significant difference in the means of frequency on a log-scale between the present study and Tomaschek, Tucker, et al. (2018), $t(40.747) = -0.690, p \approx 0.494$. The target verbs were distributed more or less normally over logarithm of frequency within and across each suffix condition, i.e., [-t] and [-(ə)n]. Because Tomaschek, Tucker, et al. (2018) found a similar lowering of the tongue tip/body for high and low frequency, and because medium frequency words showed higher tongue trajectories, we restricted our focus to high- and medium-frequency words. The ultrasound record suggests that high-frequency and low-frequency words had similar tongue shapes and positions, consistent with

the findings of Tomaschek, Tucker, et al. (2018), and are therefore not discussed further. For the purpose of visualization and ease of reference, we refer to the 10%, 50%, and 90% deciles of frequency as low, medium, and high frequency. The details of the selected items were available at `https://osf.io/3mxjg/`.

These target items was embedded with other filler items. These fillers were made of inflected forms of 213 different word types. They were either inflected verbs with their corresponding pronouns, containing other vowels than [aː], or nouns.

For the task of reading aloud these 366 target and filler word types (1017 word/phrase tokens) in total, 18 German native speakers were recruited. 10 out of the 18 participants finished all the recordings on the same day. Seven of them split the participation in the experiment to two days. One of them spent three days to complete all the recordings. The target and filler items were displayed on a laptop screen. Participants read aloud these items on the screen once for each item. While they articulated each of the items, tongue shapes and positions were recorded by ultrasound, using the software Articulate Assistant Advanced (Articulate Instruments Ltd., 2012). An ultrasound transducer was fixed under the chin of the speaker by means of a headset in such a way that the shadows created by the hyoid bone and the jaw were both visible in the ultrasound image being recorded. For a majority of the items (about 77%), ultrasound images were recorded by about 95 frames per seconds. For the other 22% of the data, 82 frames were recorded per second. For the remaining 1%, the number of frames per second ranged from 62 to 94.

The use of ultrasound in the present study, contrasting the use of EMA by Tomaschek, Tucker, et al. (2018), was motivated by their different strengths and their complementary characteristics. EMA tracks positions of sensors attached to the tongue. EMA can track some parts of the tongue very precisely, although the sensors on the tongue may hinder articulation and the very apex of the tongue is still difficult to track. This is because a sensor is placed about 5 mm posterior to

the tongue apex to avoid impeding articulations with the tongue tip. In contrast, it is very difficult to record the root of the tongue or the inside of the tongue. On the other hand, the tongue root and the inside of the tongue can be traced quite well with ultrasound, which makes ultrasound imaging a good complementary method to EMA. Accordingly, the present study made use of ultrasound imaging, aiming at replicating and extending the findings of Tomaschek, Tucker, et al. (2018) to the tongue root and the inside of the tongue.

Typically, ultrasound images of the tongue are analyzed by detecting and comparing tongue surface contours. One standard method for comparing multiple tongue surface curves is to fit spline curves to detected tongue surface positions (Davidson, 2005, 2006; Slud et al., 2002; Stone et al., 1997; Turton, 2015). Fitted splines of tongue surface contours are sometimes compared to each other by taking all the data points on the splines into consideration (Lee-Kim et al., 2013; Strycharczuk & Scobbie, 2016), while other times the data points on the splines are further summarized to representative values according to their shapes such as the curvature index (Aubin & Ménard, 2006; Bressmann et al., 2005; Noiray et al., 2013; Noiray et al., 2019).

These methods of analyzing ultrasound images all depend on detected tongue surface contours and assume the x-coordinates (horizontal positions) are accurate across frames, items, and speakers. However, this assumption does not necessarily hold true. In ultrasound midsagittal images of the tongue, the tongue tip is often hidden by shadow created by the mandible (jaw) bone, and the end of the tongue root can also be hidden by another shadow created by the hyoid bone. Because ultrasound images do not contain any anatomical reference point in themselves, it is unclear how much of the tongue tip and root is missing in actual images. In addition, tongue surface contours can flatten, e.g., for /a/, or shrink for the bulky shape required for the articulation of, e.g., /u/. Because of these characteristics of ultrasound imaging and the tongue, the same x-coordinate can actually refer to different points on the tongue surface. This issue also applies to the y-axis

coordinate, because the ultrasound transducer is usually fixed to the bottom of the chin and therefore is not consistent in its positions relative to the hard palate.

The problem of anatomical reference points can be mitigated by taking into account additional information from other parts in ultrasound images than tongue surface contour curves. To include more information from ultrasound images, several recent studies have summarized brightness values of all the pixels in ultrasound images and used these for comparison of different frames and items (Palo et al., 2014). Although it is certainly an advantage to have access to more information by including all the pixels for analysis, these methods, so far, necessarily involve summarizing ultrasound images into some representative values. Given the complexities of actual images, it may not be so straightforward to identify what is exactly different across images.

To explore the possibility to include as much information in ultrasound images as possible but not to compress pixel data into representative characteristic values, we made use of Generalized Additive Models (GAMs) and modeled brightness of each pixel in ultrasound images as a function of x- and y-coordinates, using tensor product smooths. In this approach, the tongue surface contour appears as an "area", because pixel brightness is estimated across images from different words and speakers. When there is little variability among words the predicted tongue surface contour/area becomes thin and has brighter pixel values. If there is considerable variability, the tongue surface contour/area becomes a larger area with dimmed (low) predicted pixel values. In order to take different sizes of the oral cavity into consideration, some normalization of ultrasound images is generally recommended. However, normalization does not have to be perfect. Although imprecise application of normalization would certainly lead to spurious greater variance among speakers, only a few pixels of differences in coordinates would not alter the overall results, either. In addition, such systematic differences can be accommodated with random effects in GAMMs. In what follows, we report the results obtained for the by-word ultrasound images averaged across speakers.

Comparisons with models including by-speaker random surfaces showed that averaging yields very similar mean surfaces at enormously reduced computational costs.

A downside of this statistical method is that it is computationally very demanding. In order to reduce computation time (and the concomitant carbon footprint), separate models were fitted to the data for the two suffix conditions. For each of the two suffix conditions, we fitted a model with partial effects for the x and y coordinates and their interaction, and furthermore included segment duration (`SegDur`) and (log-transformed) word frequency (`Freq`) as covariates. Segment duration was determined based on the segment boundaries obtained by a forced aligner (Rapp, 1995). We also included interactions of these covariates with the x and y coordinates and the corresponding interaction. The model specification supplied to the `bam` function of the `mgcv` package (Wood, 2017) is as follows:

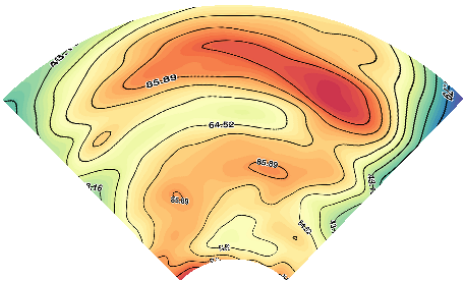
```
PixelBrightness ~ s(x, k=20) + s(y, k=20) +
                  ti(x, y, k=20) +
                  s(SegDur, k=20) +
                  ti(x, SegDur, k=c(20, 20)) +
                  ti(y, SegDur, k=c(20, 20)) +
                  ti(x, y, SegDur,k=c(20, 20, 20)) +
                  s(Freq, k=20) +
                  ti(x, Freq, k=c(20, 20)) +
                  ti(y, Freq, k=c(20, 20)) +
                  ti(x, y, Freq, k=c(20, 20, 20))
```

In order to determine the number of basis functions, we tested a range of numbers of basis functions ranging from 3 to 30. In general, estimated contours tend to be too blurred too much with lower numbers of basis functions. In the present study, numbers of basis functions being 15 or higher produced similar results. Although larger numbers of basis functions were better at capturing small details in

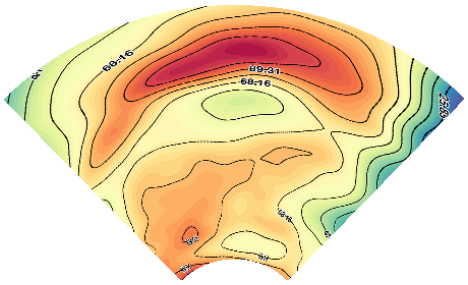
ultrasound images, greater numbers of basis functions necessarily required huge amounts of additional computation time. Therefore, we opted for 20 basis functions in the present study, aiming at a good balance between clarity of estimated contours and computational load.

The fitted models for the two suffix conditions at the middle of the vowel are visualized in Figure 2.1. For all panels of this figure, the front of the mouth is to the right. The three figures on the left side of Figure 2.1 pertain to the condition "sie...t". The three figures on the right side of Figure 2.1 pertain to the condition "sie...n". The two figures at the top visualize the ultrasound pixel brightness for high-frequency words. The two contour plots in the middle row summarize the ultrasound pixel brightness for medium-frequency words. In these top four figures, warmer colors indicate brighter pixels. The brightest pixels in the image (in dark red) are typically found close to the tongue surface. Frequency was included in the fitted models as a continuous variable. It is discretized to high- and medium-frequencies in Figure 2.1 corresponding to the quantiles 90% and 50%, to simplify visualization.

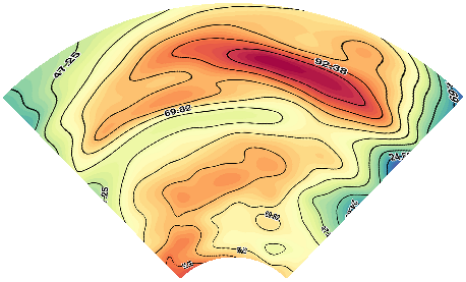
First compare the top-left panel and the top-right panel. In the condition of the suffix [t] (in the left panel), the tongue tip/blade is positioned lower and at the same time the tongue body is positioned higher, compared to the condition of the suffix [n]. Tongue fat in the tongue, indicated by the orange areas in the center of each figure below the tongue surface contours, is also pushed up in the suffix condition [t], compared to the suffix condition [n]. In addition, the tongue root is extended more posterior in the [n] condition, which is indicated by the redder area to the left of the top-right figure. The less visible tongue root in the [t] condition compared to the [n] condition indicates that the tongue root tends to be shadowed by the hyoid bone in the [t] condition, compared to the [n] condition. The forward and upward advancements of the hyoid bone are associated with the fronting of the tongue (Jordan & White, 2008; Kutzner et al., 2017; Sanders & Mu, 2013). In the current study, the forward and upward movement of the tongue is stronger



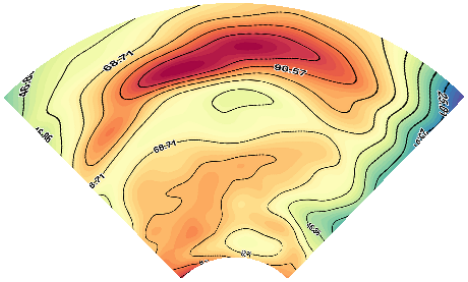
(a) High frequency words in *sie ...t*, e.g., *sie **sa**gt* [zi: z**a:**kt]



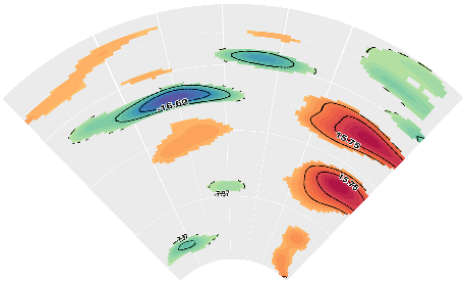
(b) High frequency words in *sie ...n*, e.g., *sie **sa**gen* [zi: z**a:**g(ə)n]



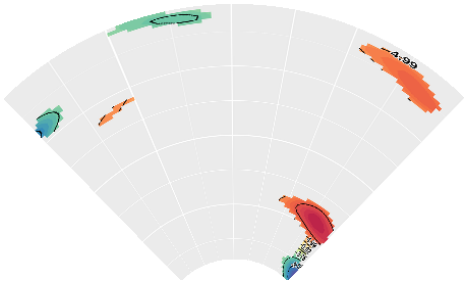
(c) Medium frequency words in *sie ...t*, e.g., *sie **na**gt* [zi: n**a:**kt]



(d) Medium frequency words in *sie ...n*, e.g., *sie **na**gen* [zi: n**a:**g(ə)n]



(e) Difference (a)-(c).



(f) Difference (b)-(d).

Figure 2.1: Predicted ultrasound images at the middle of the target vowel, i.e., [a(ː)], Warmer (darker red) colors in (a–d) are those for which the brightest pixels are predicted. Warmer colors in (e–f) represent brighter pixels in high frequency words compared to medium frequency words

in the suffix [t] condition than the suffix [n] condition. These observations hold irrespective of whether the word has a high or a medium frequency.

The two figures on the bottom present the difference surfaces between high-frequency and medium-frequency words. Warmer colors in these contour plots indicate brighter pixels for high-frequency words as compared to medium-frequency words. Colder colors, by contrast, indicate that medium-frequency words have

brighter pixels than high-frequency words. Accordingly, warmer colored regions in the bottom two figures indicate that the corresponding regions are brighter in high frequency (figures on the top), and cold colored regions indicate the opposite. The blank areas in these difference contour plots denote pixels for which there is no significant difference between high-frequency and medium-frequency words ($\alpha = 0.05$). These two difference plots are essential for clarifying where the fitted surfaces for high-frequency words and medium-frequency words actually differ.

First consider the "sie...t" condition (the left panels). In the difference plot, at the right-hand side, we find a green area (denoting brighter pixels for medium-frequency words) and below it, two red areas (denoting brighter pixels for high-frequency words). These areas represent the preferential positions of the tongue tip and tongue blade depending on the frequency of the word that is articulated. Since warmer colors represent brighter pixels in the high frequency condition compared to the medium frequency condition, the tongue tip/blade is lower for high frequency words, compared to medium frequency words.

By contrast, the left side of the difference surface (Figure 2.1e) illustrates differences pertaining to the positions of the tongue root. The orange area in the top left of the difference surface indicates a higher position of the tongue root for higher-frequency words. For medium-frequency words, as indicated by the large blue area, the tongue root is positioned lower. Below the blue area is another orange area, which we interpret to reflect tongue fat (Kim et al., 2014; S. H. Wang et al., 2020; Yu et al., 2022) that is pushed up for high-frequency words, but not for medium-frequency words.

In contrast to the suffix [t] condition, the difference surface for the suffix [n] condition (Figure 2.1f) is almost empty. Although there are some colored regions, they do not appear to be systematic. The almost empty difference surface indicates that tongue shapes are rather similar for the high- and medium-frequency conditions. The absence of an enhancement effect of frequency in the [n] condition and the presence of enhancement in the [t] condition is consistent with the findings of

Tomaschek, Tucker, et al. (2018). The absence of an enhancement effect in the [n] condition may be due to the final nasal being realized either as a syllabic nasal or as the final nasal in a separate syllable in which it is preceded by a schwa.

The present findings replicated one of the observations reported by Tomaschek, Tucker, et al. (2018), using ultrasound. The tongue tip was in a lower position for higher-frequency words as compared to medium-frequency words. This lowering of the [aː] for high-frequency words suggests these words are phonetically enhanced. The enhancement effect is in line with the hypothesis that "articulatory practice makes perfect" (Tomaschek, Tucker, et al., 2018).

In the next section, we compare inflected and non-inflected words, in order to see whether the enhancement effect observed in the ultrasound record is modulated by morphological status.

2.3 Frequency effects in relation to inflectional status

The preceding case study investigated the articulation of inflected words, with special attention to how frequency of use modulates the position of the tongue during the articulation of the stem vowel. In this section, we continue our investigation of the stem vowel, but now we impose stricter controls on segmental similarity, while comparing two conditions, one in which the stem vowel is followed by an inflectional exponent, and one in which there is no such following inflectional exponent. We call these two conditions "inflected" and "non-inflected". This allows us to investigate the consequences of inflectional status for articulation. In the literature, this contrast is often referred to as words with/without a morpheme boundary (Hayes, 2000; Lee-Kim et al., 2013). We will follow this terminology, without however assuming that some kind of boundary is physically present; i.e., the terminology will be used in a purely descriptive sense.

The goal of this second case study is to clarify whether the presence or absence of a morpheme boundary is associated with systematic differences in articulation

in EMA recordings of spontaneous conversation, under the strictest possible controls for form similarity. (Here, and in what follows, we use the term "morpheme boundary" in a descriptive sense, the presence of a morpheme boundary being equivalent to a word being morphologically complex in the concatenative sense.) It is of course impossible to find a sufficient number of word pairs with identical segments (such as German *Macht* [maxt] 'power' vs. *mach+t* [maxt] 'makes'). We therefore extracted from the Karl Eberhards Corpus of spontaneously spoken southern German (KEC) (Arnold & Tomaschek, 2016) all the words with the same rhyme structure and with the same segments for the nucleus and for the word-final segment, namely the word-final segment structure [a(ː)(C)t]. "(C)" represents at most one intervening segment between the target vowel [a(ː)] and the word-final segment [t]. *kalt* [kalt] 'cold' is one example of a non-inflected word adopted in the present study, where no morpheme boundary is located between [a(ː)] and [t]. On the other hand, *mal+t* [maːl+t] 'paints' has a morpheme boundary between [a(ː)] and [t]. The stems of the inflected and non-inflected words in our data set comprised not only simple words, but also some derived words and compounds. For example, *bemalt* consists of a prefix *be-* and a verb *malt*, where *-t* is an inflectional suffix. *Ausland* consists of a prefix *Aus-* and a noun *Land*, where no inflection is involved. In the current study, inflected words included 16 derived words by prefixation and one compound word. Non-inflected words included nine derived words and 14 compound words.

Presence and absence of a morphological boundary between the target vowel and the word-final [t] were coded manually. Intermediate cases involving stem alternation (e.g., *denken* → *gedacht*) were excluded. In the end, 84 word types, of which 48 were non-inflected and 36 were inflected, were included in the analysis, which amounts to 532 word tokens being analyzed. For each token, vertical tongue tip and body positions during the target vowel were obtained from KEC. Tongue positions were recorded in KEC with electromagnetic articulography (EMA, NDI WAVE articulograph, sample rate 400Hz). Data points outside 1.5 times the in-

terquantile range were considered to be outliers and were therefore removed from the dataset.

Vertical tongue tip/body sensor positions (`SensorPosition`) were modelled with Generalized Additive Mixed-effect Models (GAMMs) as a function of time, fitting separate models for the tongue tip and the tongue body. The sensor positions were corrected and centralized by means of three head positions and a bite plate. For details of sensors, see Arnold and Tomaschek (2016). Time was normalized, so that the onset of the vowel was 0 and the offset was 1. We included factor smooths for the interaction of speaker by normalized time.

In addition, we included random intercepts for the segments preceding and following the target vowel (`PrevSeg` and `NextSeg`). As many word types were represented by a single word token, we refrained from incorporating by-word factor smooths (Baayen & Linke, 2020) and instead included segments preceding and following the target vowel as random effects (i.e., `PrevSeg` and `NextSeg`). By including preceding and following segments, the influence of preceding and following segments on the target vowel was controlled statistically. Just in case that preceding and following segments of very frequent word tokens might distort the results, we fitted another GAMM excluding the most frequent word type (i.e., *halt* [halt] "just/simply"). This model showed very similar results as the model that included the most frequent word type. In addition, another GAMM was fitted with additional factors representing whether the segments preceding and following the vowel were alveolar, considering systematic appearance of alveolar consonants in the segments surrounding the vowel would bias the results. These additional factors did not improve the model fit and therefore were not considered any further.

We fitted separate curves for position as a function of time for non-inflected words and inflected words (`Morph`). Finally, we included duration (`SegDur`) and word frequency (`Freq`) as covariates, and allowed both covariates to interact with time and morphological complexity. Frequency was obtained from the SdeWac corpus (Faaß & Eckart, 2013) and log-transformed prior to fitting. The mean of

frequency in a log scale was higher for inflected words, compared to non-inflected words, $t(10850) = 10.541, p < 0.001$. The range of frequency was larger for inflected words, compared to non-inflected words: the maximum values of frequency were 13.084 and 12.569 for inflected and non-inflected words respectively, whereas the minimum values of frequency were 3.434 and 3.611 for inflected and non-inflected words respectively. Accordingly, there is no big difference in frequency distributions between inflected and non-inflected words, although the current data is collected from a spontaneous speech corpus (Arnold & Tomaschek, 2016) and therefore it is impossible to completely match frequency distributions between the two morphological categories. The median duration of inflected words (Mdn=0.09) was larger than the median duration of their non-inflected counterparts (Mdn=0.07, Mann-Whitney U=25004, p<0.001), which is consistent with the duration-lengthening effect found for the pre-morpheme-boundary condition (Hay, 2007; V. G. Li et al., 2020; Plag & Ben Hedia, 2018; Seyfarth et al., 2017; Smith et al., 2012; Song et al., 2013; Strycharczuk & Scobbie, 2016; Sugahara & Turk, 2009). The mean segment duration (`SegDur`) of long vowels was 121 ms and of short vowels 103 ms, $t(5231.6) = -13.866, p < 0.001$, reflecting the phonological distinction between /a/ and /a:/.

Duration and frequency were not correlated significantly when inflected and non-inflected words were aggregated, $r(510) = -0.08, p \approx 0.08$. Separating inflected and non-inflected words, non-inflected words showed a significant correlation of frequency and duration, $r(306) = -0.18, p < 0.001$, while inflected words did not, $r(202) = -0.03, p \approx 0.65$. These observations are consistent with the hypothesis that the effect of frequency on duration is modulated by morphological status.

The following GAMM was fitted to the data, again using the `bam` function of the `mgcv` package (Wood, 2017):

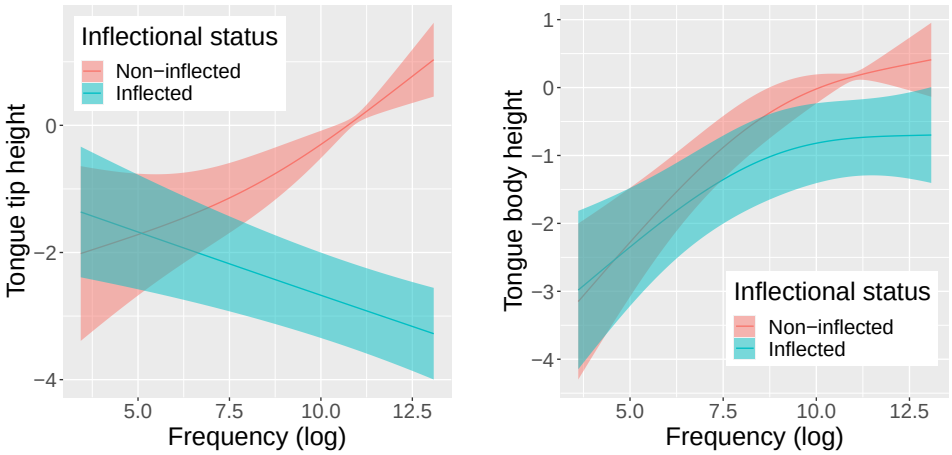
```
SensorPosition ~ s(Time, Speaker, bs=''fs'') +
                 s(PrevSeg, bs=''re'') +
```

```
s(NextSeg, bs=‘‘re’’) +
s(Time, by=Morph, k=3) +
s(SegDur, by=Morph, k=3) +
s(Freq, by=Morph, k=3) +
ti(SegDur, Time, by=Morph, k=c(3,3)) +
ti(Freq, Time, by=Morph, k=c(3,3)) +
Morph)
```

For both sensors, GAMM analyses revealed that inflected words were articulated with lower tongue trajectories on average, compared to non-inflected words, especially around the center of the vowel. In what follows, we zoom in on how frequency modulated this difference.

Figure 2.2 visualizes tongue tip/body height at the middle of the vowel as a function of frequency with inflected and non-inflected words in blue-green and red respectively. For the middle of the vowel, we selected the frame in the middle of all the frames that belong to the segment based on acoustic segmentation. For the tongue tip (left panel), we observe that for non-inflected words, a higher frequency predicts a higher vertical position. Conversely, for inflected words, a higher frequency of use predicts a lower vertical position. The right panel clarifies the effect of frequency for the tongue body sensor. The positive slope observed for the tongue tip when words are non-inflected is again present. The effect of frequency is very similar across both inflected and non-inflected words. Only for the very highest frequency words do we see a small difference between non-inflected and inflected words. Higher and lower positions correspond to reduced and enhanced articulatory realizations respectively, because, for the open low vowel [a(:)], a lower tongue height reflects lowering and hence a more clear articulation. Therefore, we observe here a reduction effect for non-inflected words both for the tongue tip and the tongue body sensor. For inflected words, an enhancement effect is observed most clearly for the tongue tip. For the tongue body, only a small difference is observed for high frequency words. This result is consistent with the earlier study

by Tomaschek, Tucker, et al. (2018), which also reported a reduced effect for the tongue body sensor.



(a) Tongue Tip

(b) Tongue Body

Figure 2.2: Vertical tongue positions (in mm) at the middle of the target vowel [a(ː)] for inflected (blue-green) and non-inflected (red) words as a function of frequency. Confidence intervals are 95% credible intervals.

Gahl and Baayen (2022) proposed that, other things being equal, the more predictable words are in utterances, the more probable it is that they will be reduced. Here, they follow the smooth signal redundancy hypothesis of Aylett and Turk (2004). At the same time, independently, again other things being equal, word-forms that are better supported by their semantics undergo articulatory enhancement (i.e., semantic strengthening). If this hypothesis is on the right track, does it follow that the present results should be understood as indicating that for non-inflected words the principle of smooth signal redundancy (Aylett & Turk, 2004) dominates (predicting higher-frequency words reduce), whereas for inflected words, the principle of semantic strengthening dominates? Of course, this immediately raises a further question, namely, why it would be only the inflected words that show an effect of semantic strengthening. This question is addressed in the next section.

In what follows, however, we first briefly consider two alternative explanations,

one building on the paradigm uniformity hypothesis, and one pursuing the smooth signal redundancy hypothesis (Aylett & Turk, 2004).

The paradigm uniformity hypothesis states that members of the same paradigm tend to be similar to each other (Hayes, 2000; Plag, 1999, 2013; Seyfarth et al., 2017). For example, Seyfarth et al. (2017) found the stem of inflected words tends to be longer than the same string of segments in morphologically simple words (e.g., ***frees*** vs. ***free****ze*). This leads to the prediction that German verb stems should also be longer compared to monomorphemic controls.

If the findings of Seyfarth et al. (2017) generalize to tongue position, one would expect to find that the vowel in stems should be articulated with lower tongue positions compared to monomorphemic controls. As can be seen in Figure 2.2, this is the case only for higher-frequency inflected words.

One reason for the absence of a main effect is that in German verb paradigms, the stem is much less dominant compared to English. Whereas in English, the stem is followed by an inflectional exponent only for the 3rd person singular, and is also used as infinitive, German verbs have somewhat richer inflection with separate forms in the singular for the three persons, and most present plurals and the infinitive sharing the same form. As a consequence, the German stem is much less dominant in its paradigm compared to English verb stems. This holds even when we take into account that the first-person singular form, which ends in a schwa, undergoes schwa-apocope in colloquial speech. Accordingly, there is no strong reason to consider that the stem of German inflected verbs is influenced strongly by their bare forms and would therefore be characterized by longer duration and less coarticulation. In fact, German inflectional verbal suffixes are alveolar, i.e., [t], [st], [(ə)n], the exception being [ə]. If anything, inflected forms with an alveolar suffix are most likely to serve as the origin of the paradigm uniformity effect.

In addition, the paradigm uniformity hypothesis does not predict an interaction with frequency. The current results show that, as frequency increases, pre-morpheme-boundary segments are articulated with tongue positions that are the

same as or lower than no-boundary segments. Therefore, the current results only partially support the paradigm uniformity account. Phonetic realizations are enhanced in complex words only when carrier-word frequency is high enough.

The paradigm uniformity hypothesis is based on the decompositional view of morphologically complex words (Taft, 1979; Taft & Forster, 1975). Strictly decompositional theories hold that inflected forms have no lexical status of their own, and are always processed on the basis of their stems, in combination with morphological rules. Therefore, these theories predict that (cumulative) stem frequency is the crucial predictor, and not the frequency of the inflected form itself. In order to evaluate this possibility, we fitted another GAM with the same structure but replaced word frequency with lemma frequency. The result showed that the fitted GAM with lemma frequency performed worse than that with word frequency (ΔML = 11.878). Therefore, the observed patterns of frequency effects according to morphological status are not very likely due to morphological decomposition and activation of lemmas.

It is also possible that places of articulation, not phonemes per se, were systematically different between morphological conditions. For example, [s] and [n] were distinguished as phonemes sharing the same place of articulation. As a consequence, it might be argued that they should undergo similar effects of coarticulation. In order to consider the possibility of potential confounding by places of articulation of segments surrounding the target vowel, we created a new factor variable that encoded whether the previous/next segment of the target vowel was alveolar. We fitted additional GAMMs with this new factor variable for tongue tip and body positions separately. In both of the models, the inclusion of the factor variable did not alter the observed effects of frequency as shown in Figure 2.2. Therefore, systematic differences in segments around the target vowel did not confound the present results regarding the observed effects of frequency.

Next, we consider how the present results challenge the smooth signal redundancy hypothesis (Aylett & Turk, 2004). The reduced realization for non-inflected

words as frequency is increased dovetails well with this hypothesis. However, this hypothesis is insensitive to morphological structure. Whether or not the target word is morphological complex, the word is predicted to be reduced if the word is likely to occur (i.e., high frequency). Consequently, this hypothesis also predicts reduction for inflected words, which is not consistent with the current results.

Since the observed interaction of frequency by inflectional status resists explanation in terms of paradigm uniformity or smooth signal redundancy, the next section investigates the consequences of inflectional semantics for articulation.

2.4 From semantics to articulation

Using the theory of the discriminative lexicon (Baayen et al., 2019), Chuang et al. (2021) and Gahl and Baayen (2022) reported phonetic enhancement for forms that are better supported by their semantics, compared to their competitors. Considering the strong form-meaning relations between inflectional suffixes and inflectional meanings, word-final forms may receive more semantic support when these forms encode inflectional semantics. This, in turn, is expected to give rise to enhanced articulation. In what follows, we pursue this explanation in two steps.

Section 2.4.1 introduces the discriminative lexicon model and formulates a measure of semantic support. Section 2.4.2 applies this measure as a covariate in a GAM model predicting vertical tongue position.

2.4.1 Deriving a semantic measure from LDL

The discriminative lexicon model (DLM: Baayen et al., 2018; Baayen et al., 2019) is a mathematical model that sets up simple mappings between numerical representations for words forms (brought together in a matrix $\boldsymbol{C}$) and numerical representations for their meanings (brought together in a matrix $\boldsymbol{S}$). These mappings are implemented using the core ideas of multivariate multiple regression, an estimation method that we refer to as Linear Discriminative Learning (LDL).

The **C**-matrix is a word by triphone matrix. For each word, the cells in its row-vector whose triphone (context sensitive phone) is contained in the word were coded as 1 and otherwise 0. The **S**-matrix brings together, for each word, a 300-dimensional vector representing its meaning, using embeddings from a pre-trained word2vec model (Müller, 2015). The $\boldsymbol{C}$ and $\boldsymbol{S}$ matrices were constructed for all the words in the CELEX database (Baayen et al., 1995) whose frequency was more than 0, and for which a word2vec (Müller, 2015) embedding was available.

A mapping $\boldsymbol{G}$ from $\boldsymbol{S}$ to $\boldsymbol{C}$ was estimated by solving $\boldsymbol{C} = \boldsymbol{SG}$. Given $\boldsymbol{G}$, a word's form vector $\hat{\boldsymbol{c}}$ is obtained by post-multiplication with $\boldsymbol{G}$ of the corresponding semantic vector: $\hat{\boldsymbol{c}} = \boldsymbol{sG}$. The sum of the semantic support for a word's triphones in $\hat{\boldsymbol{c}}$ was used by Gahl and Baayen (2022) to predict the spoken word duration of English homophones. In what follows, we focus on the semantic support for the final triphone in a word, to which we refer as `SufSemSup`, as this triphone is most relevant for the co-articulation within the vowel with the word-final [t].

Semantic support to a sublexical unit such as a word's final triphone can be understood conceptually as a measure gauging how accurately you can guess a particular sublexical string based on its carrier word's meaning. For example, the meaning "something round" does not support any particular sublexical string. Words that this meaning supports include *ball*, *apple*, *donut*, *sun*, etc. As a consequence, none of the triphones in these words receives solid support from the meaning "something round". In contrast, inflectional suffixes are usually associated well with their corresponding inflectional meanings. For example, the meaning of "past tense" predicts that it is quite likely that the word contains *-ed* at the word final position. Although the meaning of "past tense" does not exclude a few other possibilities (irregular verbs, for instance, do not inflect with *-ed*), *-ed* should receive much stronger semantic support from its inflectional meaning, in comparison to the first example with the much more specific meaning of "something round". These informal intuitions about how well a given triphone can be predicted from a word's meaning can be made precise by Linear Discriminative Learning in the

Discriminative Lexicon Model (Baayen et al., 2019).

In the preceding section, we observed an interaction of inflectional status (i.e., presence/absence of a morpheme boundary after the stem vowel) by frequency. Whereas inflectional status is a binary measure — a word is either inflected or non-inflected — `SufSemSup` is a continuous measure. To clarify whether this continuous measure is a proper real-valued counterpart for the binary predictor contrasting non-inflected and inflected words, we fitted a logistic regression model, predicting inflectional status from `SufSemSup`. Figure 2.3 visualizes how the probability of inflectional status varies with `SufSemSup`, and supports `SufSemSup` as a continuous measure of inflectional status.

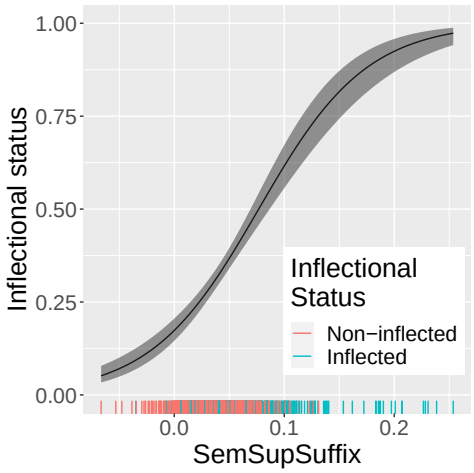


Figure 2.3: Probability of inflected words as a function of semantic support to word-final triphones. Greater semantic support for word-final triphones predicts higher probability of morphological complexity.

For the words in our dataset, we have a set of word-final triphones, all of which end with `t#`. Each of these word-final triphones comes with a different real number denoting how much support that triphone receives from the semantics of its carrier word. Importantly, the very same triphone can have different values for `SufSemSup`, as this measure critically depends on the semantic vector of its own specific carrier word. Thus, instead of having a single inflectional exponent *-t*, we have a set of final triphones with support values that depend on the word2vec

embeddings of their carrier words.

2.4.2 Predicting tongue tip positions with a measure of semantic support

Is `SufSemSup` useful as a covariate predicting vertical tongue position? And if so, how should `SufSemSup` relate to tongue tip position? In general, larger values of measures of semantic support are expected to predict more enhanced pronunciations. For instance, Gahl and Baayen (2022) and Chuang et al. (2022) observed a positive correlation of spoken word duration and measures of semantic support. For the present data, we therefore expect greater semantic support for the final triphone to be correlated with lower tongue positions, as for the [a:] vowel, lower tongue positions reflect articulatory enhancement. The reason that semantic support is expected to predict enhancement instead of reduction is straightforward. To see this, consider the form vector $\hat{\boldsymbol{c}} = \boldsymbol{sG}$. Unlike the gold standard vector $\boldsymbol{c}$, which has 1 for exactly those triphones that are part of the word, and 0 elsewhere, the $\hat{\boldsymbol{c}}$ vector has real-valued entries, which are closer to zero for triphones that are not part of a word, and that have larger positive values for triphones that are properly part of the word. The smaller the semantic support $\hat{c}_k$ for the k-th triphone is, the more likely it is that this triphone should not be articulated. Conversely, the larger the value of $\hat{c}_k$, the more likely it is that it should be pronounced. The theory of the discriminative lexicon goes one step further, and argues that the amount of semantic support correlates positively with the amount of phonetic enhancement. The above mentioned studies of spoken word duration provide support for this claim. We therefore expect that, overall, greater values of `SufSemSup` should correlate positively with phonetic enhancement.

A further question concerns whether, and if so, how, `SufSemSup` interacts with frequency. Given the results reported in the preceding section, we expect an interaction of frequency by semantic support, such that for words with low semantic support, increasing frequency predicts higher tongue positions. For words with

high semantic support, the reverse pattern is expected.

Finally, we are interested in clarifying whether SufSemSup outperforms inflectional status as predictor for the tongue tip position, as this would provide further support for the importance of taking meaning into account when studying morphological processing.

To address these questions, we fitted a slightly modified GAMM to the same dataset as analyzed in the preceding section, the only difference being that the predictor "morphological status" was replaced by SufSemSup:

```
SensorPosition ~ s(Time, Speaker, bs=''fs'') +
                 s(PrevSeg, bs=''re'') +
                 s(NextSeg, bs=''re'') +
                 s(SufSemSup, k=3) +
                 s(Freq, k=3) +
                 s(SegDur, k=3) +
                 ti(Time, SufSemSup, k=c(3,3)) +
                 ti(Time, Freq, k=c(3,3)) +
                 ti(Time, SegDur, k=c(3,3)) +
                 ti(SufSemSup, Freq, k=c(3,3)) +
                 ti(SufSemSup, SegDur, k=c(3,3)) +
                 ti(Freq, SegDur, k=c(3,3)) +
                 ti(Time, SufSemSup, Freq, k=c(3,3)))
```

All terms in this model involving frequency, semantic support, and their interaction were well supported (all $p < 0.0001$). The fitted surface spanned by frequency and SufSemSup is presented in Figure 2.4. Warmer colors indicate higher positions of the tongue tip sensor.

Except for the very lowest values of frequency, the trend is that for a fixed frequency, increasing semantic support predicts lower tongue positions. This effect is the strongest for the highest-frequency words. Thus, greater semantic support is

indeed associated with articulatory enhancement, as predicted by the theory of the discriminative lexicon.

Figure 2.4 clarifies that the expected interaction of frequency by semantic support is also present. The effect of frequency on tongue tip position has a positive slope for low values of `SufSemSup`, indicating that as frequency increases, the tongue tip sensor is found at higher positions. Since low values of `SufSemSup` indicate a low probability of being inflected, we replicate our earlier finding in the preceding section that for non-inflected words, the tongue tip rises with increasing frequency. In contrast, when word-final triphones receive good semantic support (indicated by higher values of `SufSemSup`), the slope for frequency is negative: higher probabilities of being inflected are associated with lower tongue positions.

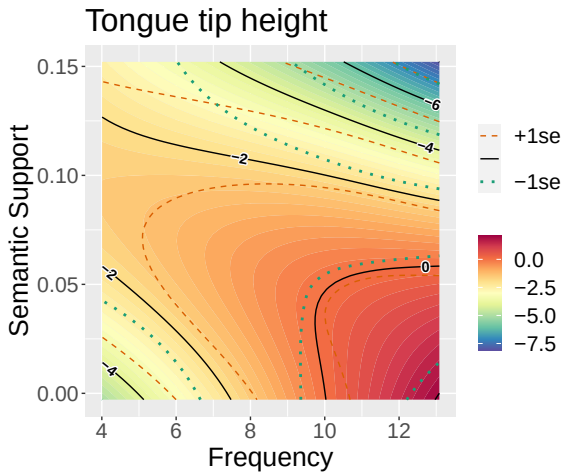


Figure 2.4: Interaction of effects of frequency and `SufSemSup` at the middle of the vowel [a(ː)]. Warmer and colder colors represent higher (reduced) and lower (enhanced) tongue tip positions respectively. Dashed lines specify 1SE confidence regions for the contour lines.

This interaction between frequency and `SufSemSup` is not confounded with triphone frequency. Although `SufSemSep` is mildly correlated with triphone frequency, $r(72) = 0.36$, $p < 0.01$, adding triphone frequency to the GAMM does not eliminate the effect of `SufSemSup`. As `SufSemSup` increases, the tongue tip sensor is positioned lower. By contrast, an increase in triphone frequency is associated

with tongue raising. Within the framework of DLM (Baayen et al., 2019), the effect of triphone token frequency is likely an effect of triphone type frequency in disguise. Triphones that occur in more different words and that realize more different senses are more difficult to learn, and hence will receive reduced support from their senses. A new algorithm for incorporating frequency of use into the estimation of the mapping from meaning to form is expected to allow for more precise modelling of the semantic support for the final triphone while also taking into account the consequences of the frequencies with which triphones occur (Heitmeier et al., 2022).

Similarly, the observed patterns do not seem to be confounded with the transitional probabilities of phonemes into the target vowel. To look into the potential confound by transitional probability, we fitted another GAMM by adding bigram conditional probability of the stem vowel given one segment before the vowel to the GAM model. The inclusion of transitional probability, however, led to worse model performance in spite of increased model complexity ($\Delta\text{ML} = 13.065$). Therefore, transitional probability into the target vowel is not an essential predictor for the present dataset.

A remarkable result is that replacing the original categorical predictor "inflectional status" with `SufSemSup` results in a substantial improvement in model fit by no less than 10158 AIC units (model comparison test: $\chi^2(5) = 5051.669$, $p < 0.0001$). Importantly, at no point is morphological structure explicitly coded into the computational model: form representations are based on triphones, and semantic representations are based on word2vec, which, unlike fasttext, has no access to internal word structure. All that is required is inspection of the word-specific semantic support for the final triphone. Theoretical constructs such as morphemes and morpheme-boundaries are not required.

Many different semantic measures can be derived within the framework of the DLM, in addition to semantic support. In fact, eight other semantic measures were actually computed and evaluated for the present study. However, based on variable

importance estimated with a random forest, it was the `SufSemSup` that turned out to be the most powerful predictor of tongue tip height (although several other related measures also performed quite well). The importance of specifically this final triphone may help explain why there is no strong frequency effect for the tongue body sensor: the tongue body is less involved in the co-articulation with the word-final [t]. Slightly higher tongue body positions found in the first study with ultrasound were therefore likely due to a passive movement induced by the lowering of the tongue tip. In addition, it is possible that what we referred to as "tongue body" in ultrasound images can correspond to somewhere more in the back of the oral cavity than the location of the "tongue body" sensor of EMA. Because of a lack of anatomical reference points in ultrasound imaging, it is far from straightforward to establish how the positions of the EMA sensors correspond to regions in the ultrasound image. This topic is therefore left for future research using simultaneous recordings with EMA and ultrasound.

2.5 Discussion

In this chapter, we presented three case studies. The first study replicated one of the previous findings reported by Tomaschek, Tucker, et al. (2018), using ultrasound. Tomaschek, Tucker, et al. (2018), using EMA, found enhanced articulatory realizations for high frequency inflected words, compared to medium frequency words. A similar enhancement effect emerged from the ultrasound recordings. The more enhanced realization of higher frequency words supports the hypothesis of kinematic improvement with practice.

In the second study, we compared articulatory realizations of [a(ː)] of inflected and non-inflected words. For non-inflected words, a higher frequency predicted a higher tongue position, suggesting articulatory reduction. But for inflected words, the opposite effect emerged, most notably for the tongue tip sensor, which is more actively involved in the co-articulation between [a(ː)] and [t].

This interaction between frequency and inflectional status is not predicted by the paradigm uniformity hypothesis (Seyfarth et al., 2017), the smooth signal redundancy hypothesis (Aylett & Turk, 2004), the paradigmatic signal enhancement hypothesis (Kuperman et al., 2007), and the kinematic practice hypothesis (Tomaschek, Arnold, et al., 2018; Tomaschek, Tucker, et al., 2018). All these hypotheses can explain either the pattern for inflected words, or the pattern for non-inflected words, but not both.

Following up on studies that investigated how semantic support for forms of words affects spoken word duration (Chuang et al., 2021; Gahl & Baayen, 2022), the third study replaced the categorical predictor of inflectional status (inflected vs. non-inflected) by a continuous measure of semantic support provided by empirical word embeddings (word2vec) for the word-final triphone. This measure of semantic support was shown to be the real-valued counterpart of the categorical measure: the probability of being inflected increases with semantic support. Replacing morphological status with semantic support resulted in a substantially improved model fit. For words with low semantic support (most likely non-inflected words), tongue height was positively correlated with frequency; for words with high semantic support, the correlation changed sign. This result indicates that it is the specific semantics of inflected and non-inflected words (e.g., inflectional meanings) that drive articulation, and not the presence of an exponent or a putative morpheme boundary in forms of words. In other words, systematic differences in meanings create different (co-)articulatory patterns between inflected and non-inflected words, and not as a consequence of the presence of a discrete exponent or hypothesized morpheme boundary. Systematic differences in meaning are not restricted to inflected and non-inflected words. Investigation of the semantic support for derived words is a topic that is on our agenda for future research.

The observed reduction and enhancement effects of frequency according to different degrees of semantic support to inflectional suffixes can also be understood to elaborate the distinction between the need for clearer articulation in favor of

the listener and the desire to save articulatory effort for the speaker themselves (Lindblom, 1983; Lindblom & Marchal, 1990; Nelson, 1983).

In general, it takes longer to move the tongue across longer distance (Kelso et al., 1985). In order to keep the same "traveling" duration, the speaker has to move the tongue faster. For moving the tongue faster, the speaker needs to make greater articulatory effort (Nelson, 1983). According to the H&H theory (Lindblom & Marchal, 1990), the speaker has certain degrees of freedom with respect to articulatory effort. They can choose to make more effort to move the tongue for clearer articulation under temporal pressure, whereas they can also choose to undershoot an articulatory target (less clearer, reduced articulation) for saving articulatory effort (Lindblom, 1983). Lindblom and Marchal (1990) suggested that the balance of the benefits for the speaker and the listener is determined by various communicative factors. Therefore, although duration is correlated with longer distance and faster tongue movements, duration cannot be a sole factor to determine degrees of articulatory reduction/enhancement (Lindblom, 1983; Lindblom & Marchal, 1990).

The present study controlled vowel duration and phonological environment. And yet we found clearer articulation for the words of higher frequency and greater semantic support. Therefore, the present findings can also be interpreted to elaborate other factors than just duration, which are responsible for the many low-level phonetic differences which Lindblom and Marchal (1990) called "the lack of invariance" (Lindblom & Marchal, 1990, p. 403). More precisely, the present study suggests that, when duration is kept constant, the speaker tends to pay more articulatory effort for clearer (co-)articulation when the suffix of the word is semantically motivated and therefore less uncertainty is involved for the phonetic makeup of that word.

More direct evaluation of the claims by the H&H theory (Lindblom & Marchal, 1990) is possible. A number of studies has suggested that the balance of the benefits for the speaker and the listener is important to predict phonetic realization

(Kelso et al., 1985; Lindblom, 1983; Lindblom & Marchal, 1990; Nelson, 1983). The present study focused on a semantic measure purely from the perspective of speech production. However, the counterpart of `SufSemSup` on the comprehension side, called functional load, was also found to be predictive for phonetic realization (Denistia & Baayen, 2022; Saito et al., 2021). `SufSemSup` and functional load are expected to capture opposing forces between the benefits for the speaker and the listener, which are analogous to the distinction of the clarity for the listener and the economy of articulatory effort according to Lindblom and Marchal (1990). This possibility is left open for future research.

Another open question is how to understand the observed effect of frequency. The current results only indicate that the balance between the reduction and enhancement effects of frequency are adjusted by semantic support for forms. It is perhaps unsurprising that the effect of semantic support is strongest for the highest-frequency words, which are the words for which we have more observations, and that speakers have encountered more often. However, this role of frequency for learning is not properly accounted for by the way in which we estimated the mapping from meaning to form, namely, using the matrix algebra of multivariate multiple regression. The resulting mapping represents the "endstate of learning" that is reached with "infinite" experience, see Heitmeier et al. (2021) and Shafaei-Bajestan et al. (2021) for detailed discussion. Recently, a new algorithm has been developed that is able to take frequency of use into account (Heitmeier et al., 2022); Taking frequency of use into account for the mapping between meaning and form then is another topic that is high on our research agenda.

How frequency shapes the fitted surface for tongue tip height likely reflects two factors. The first factor concerns how frequency modulates the learning of the mapping from meaning to form. The present estimated form vectors $\hat{c}$ are suboptimal in this respect, and hopefully can be improved in the near future.

The second factor concerns how the informativity of the word in the discourse (Aylett & Turk, 2004) affects articulation. Within the framework of the discrim-

inative lexicon, given a measure of the informativity $h_{\omega,k} \geq 0$ of a word token ω at point k in a discourse, its effect on the semantic support for triphones is, in the simplest possible scenario, proportional to $h_{\omega,k}\hat{c}$. In this way, the reduction effect and the enhancement effect of frequency can be represented jointly in the model. However, teasing apart the independent contributions of these two factors to articulation remains a topic for further research. Whether the paradigmatic enhancement hypothesis (Kuperman et al., 2007) can be integrated within the present approach likewise awaits further advances in computational modeling.

What the present study is able to contribute to the advancement of knowledge is, first, further support for the possibility that part of the effect of frequency may reflect articulatory practice, and second, that quantitative representations of meanings of words, such as made available by distributional semantics, can be used to obtain substantially more precise predictions of phonetic realization.

Chapter 3

Analyzing ultrasound images with GA(M)Ms

Abstract: This study has two aims: 1) to develop a new analysis methodology of ultrasound images, and 2) to investigate effects of phonetic enhancement by word frequency, using ultrasound. The new analysis method of ultrasound images makes use of Generalized Additive Mixed-effect Models (GAMM: Wood, 2017) and predicts each pixel brightness to constitute whole ultrasound images. Using this new methodology, in the following theoretical part of this paper, we address the phonetic enhancement effects of word frequency with experimental data of German native speakers articulating pronoun-verb combinations. The results indicated that more peripheral tongue positions for the stem vowel [a(ː)] were associated with higher frequency words. This effect of word frequency was observed for the suffix [t] condition, but not for the suffix [n] condition. These results will be interpreted to support general effects of phonetic enhancement by high frequency, explained by the paradigmatic signal enhancement hypothesis (Kuperman et al., 2007) and the kinematic practice hypothesis (Tomaschek, Arnold, et al., 2018; Tomaschek, Tucker, et al., 2018).

3.1 Introduction

Tongue movements during speech can provide extensive insights to speech production process. To record tongue shapes and movements during speech, several different techniques have been developed. One of these techniques is ultrasound imaging. Ultrasound imaging makes use of high frequency sound waves that in phonetics are used to trace the tongue surface.

Although ultrasound imaging has attracted more and more attention and has been adopted in a wide range of studies, there is still no consensus regarding the optimal way for analyzing ultrasound images and sequences of ultrasound images. A standard methodology fits a spline curve to the brightest pixels that are expected to correspond to the tongue surface contour. This methodology has the advantage that it reduces the high complexity of ultrasound images only to just a set of pixel locations that are straightforwardly described by their x and y coordinates. Detected tongue contours can then be straightforwardly compared with statistical methods such as Smooth Spline ANOVA (Davidson, 2006; Gu, 2002; Lee-Kim et al., 2013; Strycharczuk & Scobbie, 2016; Sung, 2014) or GAMs (Generalized Additive Mixed-effects Models: Heyne et al., 2019; Noiray et al., 2019; Strycharczuk & Scobbie, 2017).

However, extraction of only the tongue surface contour has the disadvantage that other potentially relevant information about tongue movements are not available to the analyst. For example, changes in the location of tongue fat within the tongue can be indicative of changes in muscle tension and contraction taking place during articulation. More importantly, movements of the hyoid bone are visible in the ultrasound record (Hiiemae et al., 2002; Ma & Wrench, 2022; Rossi & Autesserre, 1981) and can inform about tongue fronting and tongue elevation, although there is still a lot of uncertainty as to the exact interpretation of the data (Buchaillard et al., 2009).

This study proposes to analyze full ultrasound images with the Generalized Additive Model (GAM: Wood, 2017), using tensor product smooths. Tensor prod-

uct smooths can take into account all the pixels in an ultrasound image, and the predicted pixel brightness can be visualized using contour plots.

In what follows, we first summarize advantages and disadvantages of ultrasound imaging in relation to other tongue-shape-recording techniques (Section 3.2.1). Subsequently, we introduce several key muscles in and around the tongue (Sections 3.2.3-3.2.5), as a basic understanding of tongue muscles is necessary for properly interpreting ultrasound images. Next, we briefly explain how ultrasound imaging works (Section 3.2.6), and then discuss existing methods for analyzing ultrasound images (Section 3.2.2). Against this background information, we then show how tensor product smooths can be used to analyze full ultrasound images (Section 3.3). Section 3.4 illustrates the method for experimental data on the production of the German [a(ː)] vowel.

3.2 Background

3.2.1 Why ultrasound?

The present paper introduces a new method of analyzing ultrasound images. In the context of articulatory phonetics, ultrasound imaging is one of the tools for recording the tongue shape or position during speech. Other tools are X-ray photography (Liljencrants, 1971), Cinefluography (Harshman et al., 1977), X-ray microbeam registration (Westbury, 1994), Electromagnetic Articulography (EMA) (Arnold & Tomaschek, 2016; Cho, 2001; Dang et al., 2008; Dang et al., 2009; Erickson et al., 2014; Hertrich & Ackermann, 2000; Lee et al., 2019; Perkell et al., 1992; Saito et al., 2021; Steele & van Lieshout, 2004; Tiede et al., 2011; Tomaschek, Arnold, et al., 2018; Tomaschek, Tucker, & Baayen, 2019; Tomaschek, Tucker, et al., 2018; Tomaschek et al., 2013; J. Wang et al., 2013), and Magnetic Resonance Imaging (MRI) (Masaki et al., 1996; Moisik et al., 2019; Stone et al., 2001).

The oldest technique for tracing tongue position made use of X-rays. X-ray photography takes a few snapshots of the oral cavity (Liljencrants, 1971). Cine-

fluography takes a series of X-ray photographs, which are combined to construct a video capturing tongue movements (Harshman et al., 1977). X-ray photography is not well-suited to record movements of the tongue. Although cinefluography can take a video of tongue movements, extended radiation exposure comes with a potential health risk. X-ray microbeam registration dealt with this problem by attaching gold pellets to the tongue and by focusing a microbeam specifically on these pellets (Westbury, 1994). As a consequence, only small parts of the oral cavity are exposed to radiation. However, only the positions of the pellets are traced. Furthermore, this method suffered from limited accessibility (Steele & van Lieshout, 2004).

X-ray microbeam registration is a point-tracking system. Another point-tracking system, electromagnetic articulography (EMA), is more accessible and widely used. Similarly to X-ray microbeam registration, EMA requires several sensors to be attached on the tongue and the lips. The EMA system creates an electromagnetic field, which induces current in sensors. The changes in the current strength are used to estimate positions of sensors. Since the amount of current produced stands in an inverse relation to the cube of the distance, the precise locations of the sensors can be calculated. Disadvantages of EMA are 1) that sensors need to be glued to the tongue and the lips, which is practically impossible to do for the back and root of the tongue, due to the gag reflex, 2) that only the positions of the sensors are traceable, not the entire tongue surface, 3) that sensors are difficult to place at exactly same positions across experiments with the same speaker, and within an experiment when sensors get detached, and 4) that sensors and wires connecting sensors to the EMA system may interfere with natural articulation.

These disadvantages are absent when ultrasound or MRI are used. Ultrasound and MRI both capture a full image of the tongue. Ultrasound makes use of high frequency sound waves and MRI uses a strong magnetic field. While MRI provides clearer pictures of the tongue than ultrasound, it has several downsides. One of its disadvantages is its low time resolution, which can be a problem for recording rapid

movements of the articulators. Furthermore, the speaker has to lie down when they are recorded. Different body postures come with different ways in which gravity affects tongue positions and tongue shapes (Hoedl, 2015). However, it is worth noting that these downsides may perhaps be solved in the near future thanks to the development of real time MRI systems (Nayak et al., 2022) and upright MRI registration (Botchu et al., 2018). Practical disadvantages of MRI are 1) that MRI machines are expensive both with respect to acquisition and maintenance, 2) that MRI machines are not portable and hence not practical to use for studying articulation in children, and 3) that MRI registration is not without health risks and cannot be used with participants with implants such as pacemakers.

Ultrasound systems are relatively cheap and transportable. Because of their better mobility, ultrasound systems can also be used for field work (Gick, 2002). Ultrasound images can be recorded with good temporal resolution, they capture the whole tongue, they are not invasive, and don't come with health risks. These characteristics make ultrasound an especially suitable tool for children (Noiray et al., 2013; Noiray et al., 2019).

These advantages of ultrasound come at the cost of less visual clarity. Ultrasound images usually contain a lot of noise and possibly visual artefacts. In addition, ultrasound images usually do not contain clear anatomical reference points. Furthermore, a time-series of ultrasound images poses challenges for interpretation due to both movements of the transducer with respect to the vocal tract, and movements within the vocal tract that change the locations of anatomical reference points. These problems also arise when the differently shaped vocal tracts of individual speakers have to be taken into account. As a consequence, analyzing ultrasound images can be challenging, an issue to which we will return below.

3.2.2 Analysis methods for ultrasound images

While ultrasound imaging has attracted more and more attentions for phonetic, phonological, and also clinical purposes, there is no consensus about how to best

analyze ultrasound images. A majority of studies has focused on tongue surface contours. Some analysts zoomed in on representative values of these contours, for example, the x- and y-coordinates of the highest point on the tongue surface contour (Lin et al., 2011; Noiray et al., 2013; Noiray et al., 2019). Other analysts have taken more points of the tongue surface contour into consideration. In their approaches, the shape of the tongue surface contour is mapped onto one or more representative values (Aubin & Ménard, 2006; Bressmann et al., 2005; Davidson, 2005; Dawson et al., 2016; Ménard et al., 2013; Slud et al., 2002; Song et al., 2013; Stolar & Gick, 2013; Stone, 2005; Stone et al., 1992; Turton, 2015). One of the simplest methods is to approximate the shape of the tongue contour with a triangle (Aubin & Ménard, 2006; Ménard et al., 2013; Song et al., 2013). The base of the triangle is set up by connecting both ends of the tongue surface contour. The height of the triangle is given by the longest perpendicular line starting from the base and ending at the tongue contour. If this perpendicular line is longer than the base line, then the tongue is more bunched up. In contrast, if the base is longer than the perpendicular line, the tongue is more flattened. Measures based on several ratios calculated from the triangle were used by Aubin and Ménard (2006) to study compensatory tongue movements in adults and children, by Ménard et al. (2013) to investigate differences between blind and sighted speakers, and by Song et al. (2013) to trace different articulatory realizations of consonant clusters according to their morphological properties.

The curvature index (CI) and the modified curvature index (MCI) are more sophisticated measures for assessing the shape of the tongue surface contour (Dawson et al., 2016; Stolar & Gick, 2013). The curvature index is defined as the integral of the radius of curvature from one end of the tongue contour curve to the other, which is fitted with a 7th-order polynomial function (Stolar & Gick, 2013). Dawson et al. (2016) introduced a slightly modified version of this index (MCI), the differences being 1) that the integral was based on the arc length rather than the x-coordinates, and 2) that central differencing replaced polynomial based differen-

tiation. The CI and the MCI both have the advantage of interpretability. Higher (modified) curvature index values indicate greater curvature with greater complexity in the tongue contour line. In addition, these measures are robust to rotations of the tongue surface contour. Such rotations can occur due to movement of the ultrasound transducer, one of the problems inherent to ultrasound imaging. In this respect, the MCI outperformed the CI. For the formal definitions of the curvature indices, see Appendix 3.A.

Similarly to the CI and the MCI, the so-called Procrustes distance (J. Wang et al., 2013) produces a single value representing the complexity of the tongue surface contour. A Procrustes analysis is a statistical method that measures the distance between two shapes in such a way that the distance between the points of the two shapes is minimized. This is accomplished by aligning, scaling, and rotating. The distance between the aligned shapes is defined as the sum of Euclidean distances between the points (landmarks) of the two shapes. To apply a Procrustes analysis to the tongue surface contour, the tongue shape in resting position is used as the reference shape. The distance from this reference shape to the target tongue shape during articulation is the Procrustes distance (Dawson et al., 2016; J. Wang et al., 2013).

While the Procrustes distance and the MCI provide interpretable measures of curvature, they have been found to be less effective for delineating phonemes, compared to discrete Fourier Transform (DFT) (Dawson et al., 2016). A DFT maps a sequence of equally-spaced samples of a function represented in the time domain to corresponding data points in the frequency domain. In other words, DFT looks for sinusoid functions with different frequencies, such that the weighted sum of the sinusoid functions converges to the shape function. For more details, see Appendix 3.B. The weights of the sines are the output of the DFT method and form the input for the analysis of the tongue surface. Dawson et al. (2016) used the DFT, the MCI, and the Procrustes distance to predict phoneme identity from the tongue surface contour, and reported that the DFT outperformed the MCI and the Procrustes

distance in terms of classification rates by Linear Discriminant Analysis (LDA).

The original work that applied Fourier series to trace the tongue surface contour was Liljencrants (1971). Liljencrants (1971) adopted the real-value counterpart of DFT. These authors approximated the tongue contour by a sum of cosine and sine functions with different coefficients. They reconstructed the tongue contour from only the first few Fourier coefficients using the inverse transform. The differences between the original tongue contour and the reconstructed tongue contour were evaluated in terms of the root mean square (RMS). According to Liljencrants (1971), the DC component (i.e., zero frequency component) and the fundamental frequency already capture the overall shapes of tongue contours quite well. By adding the second harmonic, reconstructed tongue contours become very similar to the originals.

One downside of Fourier based analysis is that coefficients are far from straightforward to interpret. Although Dawson et al. (2016) found that the imaginary part of the first DFT coefficients discriminated between phonemes, it was not clear what the imaginary part of the first DFT actually represented. In addition, the magnitude of the DFT coefficients did not correspond to the degree of phoneme complexity. In Dawson et al. (2016), phonemes were categorized in terms of the complexity of their tongue shapes, as low, middle, or high complexity. For example, /ʌ/ was categorize as low complexity, because it does not require movements of the tongue (as for glides) or second constriction (as for /ɹ/). For example, a lower value of the imaginary part of the first DFT coefficient was associated with a high phoneme complexity, a higher value was associated with medium complexity, and a middle value was associated with low complexity.

The methods discussed thus far all involve compressing or transforming coordinates of the tongue contour into some representative values. Other studies have compared multiple tongue contours without compressing or transforming tongue surface coordinates (Davidson, 2006; Heyne et al., 2019; Lee-Kim et al., 2013; Strycharczuk & Scobbie, 2016). Davidson (2006) made use of Smooth-

ing Spline ANOVA (Gu, 2002), which makes it possible to evaluate statistically whether tongue contours linked to experimental factorial manipulations are significantly different. Inferred tongue contours can be visually represented together with their confidence intervals. Davidson (2006) used this statistical method to study coarticulation coarticulation of English /g/ in word-final position. Lee-Kim et al. (2013) used Smoothing Spline ANOVA to investigate the darkness of English /l/ before and after morpheme boundaries. A downside of this method is that it is very sensitive to small systematic differences in tongue contours that are linked to the geometry of individual speakers' vocal tracts. Therefore, Smoothing Spline ANOVA is not recommended for multiple tongue contours from different sessions and speakers.

Recently, Generalized Additive Mixed Models (GAMMs) (Wood, 2017) have been employed to compare multiple tongue contours (Heyne et al., 2019). GAMMs are an extension of the standard linear regression model that can estimate not only linear effects but also non-linear effects of predictors. Just as Smoothing Spline ANOVA, smoothing splines play a central role in GAMMs. But GAMMs make many different kinds of splines available for both single predictors and multiple predictors. Heyne et al. (2019) recorded the tongue shapes of English and Tongan speakers during articulating vowels using ultrasound. Tongue contours were traced manually by placing several points on the tongue surface contour and by interpolating these points with 100 points of a cubic spline. Tongue height was then defined as the distance from the origin to the tongue surface contour. Tongue height and variance in tongue height were modeled as a function of a range of predictors including random effect smooths for speakers. Heyne et al. (2019) showed that there was more variability in the way Tongan vowels were produced by native speakers of Tongan, compared to English vowels produced by native speakers of New Zealand English.

Although Smoothing Spline ANOVA and GAMMs extended the analysis of ultrasound images from representative values to full tongue contours, these methods

make use of only a part of the information contained in ultrasound images. Furthermore, ultrasound images can contain a lot of noise, which leads to uncertainty about the exact position of the tongue surface. As a consequence, any analysis done on an imputed tongue surface inherits the errors made by the tongue surface tracker. Even when tongue tracking is accurate, it is often the case that the full tongue contour cannot be traced. Because ultrasound pulses are blocked by air or bones, ultrasound images often have big shadows around the tongue tip and the tongue root, rendering these invisible. As a result, the parts of the tongue contour that are visible can differ in length. To deal with such different lengths, normalization is usually carried out (Heyne et al., 2019; Liljencrants, 1971; Slud et al., 2002; Stone et al., 1992; Stone et al., 1997; Strycharczuk & Scobbie, 2016). For example, shorter tongue contours can be stretched to match the longest tongue contour (Slud et al., 2002). However, stretching comes with the risk of comparing different parts of the tongue surface. For example, when the tongue tip is raised for articulating a /t/, a large part of the tongue tip becomes invisible. In contrast, when the tongue tip is in rest with no air pocket underneath it, as when producing an /a/, more of the tongue tip is visible. When these two tongue surfaces are compared after normalization, not taking into account how much of the tongue tip is hidden by the mandible shadow, will lead to equation of the front part of the tongue tip with the rear part of the tongue tip or even the tongue blade.

This issue can be mitigated by including information provided by other sources than the tongue contour such as the hyoid shadow. To include more information than the tongue contour, Palo et al. (2014) introduced the Pixel Difference (PD) method. This method calculates mean Euclidean distances across all the pixels between frames (Palo, 2019, 2020, 2021; Palo & Lulich, 2021; Palo et al., 2014), making it suitable for tracing changes in ultrasound images over time. Using this pixel difference method, Palo et al. (2014) reported that the initiation of articulation can precede the onset of audible speech. A disadvantage of this method is that it is difficult to establish where differences in pixel locations between ultrasound

images are significant.

Fortunately, the generalized additive model (GAM Wood, 2017) offers flexible statistical tools to model full ultrasound images and their development over time. For the interpretation of full ultrasound images, it is necessary to have some understanding of the anatomy of the tongue.

3.2.3 Tongue muscles

Anatomical descriptions of the tongue distinguish between intrinsic muscles, extrinsic muscles, and neck muscles. In what follows, we discuss the sets of muscles in further detail.

Intrinsic muscles

The inside of the tongue is made up of four muscles: the superior longitudinal muscle, the inferior longitudinal muscle, the transverse muscle, and the vertical muscle (Figure 3.1).

The superior longitudinal muscle extends from the middle line of the tongue (the median septum) and stretches out to the edges of the tongue including the tongue apex. This muscle is used to raise the tongue tip for alveolars and to curl up for retroflex consonants. In addition, by contracting the superior longitudinal muscle, the tongue root can be slightly pushed out backward (Buchaillard et al., 2009).

The inferior longitudinal muscle stretches from the hyoid bone and its surroundings to the inferior part of the tongue tip. The hyoid bone is a small U-shape bone located at the bottom of the tongue root. Since the inferior longitudinal muscle is located in the inferior part of the tongue, its contraction causes the tongue tip to be pulled downward and backward. This lowering action is required for the release of tongue tip for stop consonants. In addition, by lowering and retracting the tongue tip/blade, the tongue body can be bulged, which is required for articulation of velar consonants (Epstein et al., 2002). The inferior longitudinal muscle

also helps to shorten and stiffen the tongue, working together with the superior longitudinal muscle.

The transverse muscle runs from the median septum of the tongue towards the edges of the tongue. It is located below the superior longitudinal muscle. Its function is mainly to narrow the tongue (Shaw & Martino, 2013). Narrowing the tongue at the same time elongates the tongue. This elongation can be observed as a small forward movement of the tongue tip and/or a small backward movement of the tongue root (Buchaillard et al., 2009). As a consequence, the transverse muscle is involved in the articulation of front vowels and consonants (Epstein et al., 2002).

The vertical muscle stretches down from below the superior longitudinal muscle to the inferior longitudinal muscle. Its vertical fibers are interlaced with the transverse fibers of the transverse muscle. By contracting, the vertical muscle flattens and widens the tongue. Flattening and widening cause the tongue to approach and contact the (hard) palate, which is required for high front vowels and alveolar stops (Epstein et al., 2002).

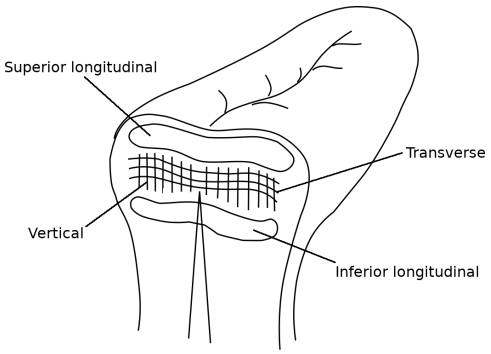


Figure 3.1: Schematic image of intrinsic muscles.

Extrinsic muscles

The intrinsic muscles make up the top part of the tongue. This top part of the tongue is supported externally by several other muscles, the so-called extrinsic muscles. There are four extrinsic muscles: the genioglossus, the hyoglossus, the styloglossus, and the palatoglossus (Figure 3.2). The names of these muscles indicate where

these muscles originate.

The genioglossus (*genio-*, "chin", *-glossus*, "tongue") makes up the main part of the inner tongue. This muscle extends from the superior mental spine, which is located in the center of the mandible (jaw), to the dorsum of the tongue and also to the hyoid bone. This muscle extends in a fan-shaped way to the entire surface of the tongue, including the tongue root. This versatile muscle is a major player in articulation. The anterior part of the genioglossus extends to the tongue tip. Its contraction pulls the tongue tip downward and backward. This also causes the tongue root to be slightly pushed out backward (Buchaillard et al., 2009). The movement of the tongue tip is involved in the release action of alveolar stop consonants. Important for the interpretation of ultrasound images is that contraction of the anterior genioglossus elevates the hyoid bone (Epstein et al., 2002). The medial genioglossus is involved in lowering the tongue body, which also causes the tongue tip to be pushed out, and slightly up (Buchaillard et al., 2009). The posterior genioglossus muscle extends all the way to the tongue root. When contracting, it pulls the tongue root forward. This in turn causes the upper part of the tongue to be pushed up and forward, resulting in raising of the tongue tip and the tongue blade (Buchaillard et al., 2009). Thus, the posterior genioglossus is also involved in the articulation for many sounds made in the front of the mouth (Epstein et al., 2002).

Both sides of the tongue are supported by the hyoglossus, which connects the hyoid bone (*hyo-*) and the tongue (*-glossus*). The hyoid bone is located in the tongue root. Contraction of the hyoglossus pulls the tongue root backward and the tongue body downward (Buchaillard et al., 2009). While the main functions of the hyoglossus are lowering and retraction of the tongue, it also balances the forward movement of the whole tongue controlled in part by the posterior genioglossus. For front vowels and back vowels, the hyoglossus acts as antagonist[1] to the styloglossus (Epstein et al., 2002). The styloglossus lifts up the back of the tongue, whereas the

[1] Antagonist muscles are muscles that exert a force opposite to that of primary, agonist muscles.

hyoglossus muscle antagonistically helps to control the position of the tongue back by pulling it down and retracting it at the same time.

The styloglossus is a muscle that attaches to the back part of the tongue and to the temporal styloid process. The temporal styloid process is a small cylindrical bone located right beneath the ears. Contraction of the styloglossus elevates and retracts the tongue body. The styloglossus is essential for articulation of most back vowels (Epstein et al., 2002). In addition, the styloglossus pulls the tongue tip downward and backward, probably as a mechanical consequence of tongue body elevation and retraction (Buchaillard et al., 2009).

Elevation of the tongue body by the styloglossus is further supported by the palatoglossus. The palatoglossus is a muscle connecting the end of the hard palate with the side of the back of the tongue. Together with the styloglossus, the palatoglossus contributes raising and buldging the tongue body (Shaw & Martino, 2013). This muscle is required for velar consonants (Epstein et al., 2002).

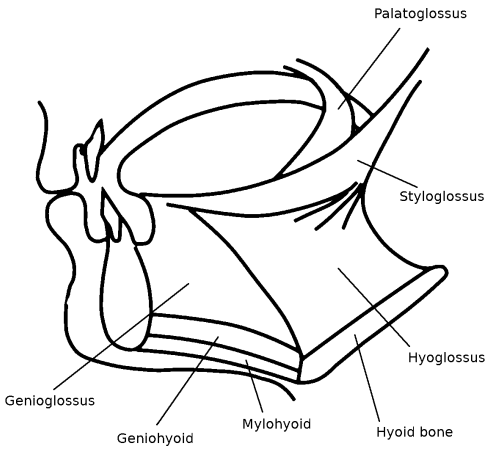


Figure 3.2: Schematic image of extrinsic and neck muscles.

3.2.4 Neck muscles

In addition to the extrinsic and intrinsic muscles for the tongue, ultrasound captures two additional muscles: the mylohyoid and the geniohyoid (Figure 3.2). The mylohyoid is the lowest muscle in the mouth and connects the inner lowest part of

the mandible bone and the hyoid bone. It is often described as the floor muscle of the mouth. Its contraction pulls the hyoid bone forward and upward (Epstein et al., 2002; Shaw & Martino, 2013). At the same time, the floor of the mouth is elevated when the mylohyoid contacts, causing elevation of the tongue body (Buchaillard et al., 2009). The mylohyoid contributes to the articulation of velars and high back vowels. (Epstein et al., 2002). During ultrasound registration, the transducer is placed below the chin and hence is directly next to the mylohyoid.

The geniohyoid is located directly above the mylohyoid and connects the lower central part of the mandible with the hyoid bone. This muscle is more tubular in shape, whereas the mylohyoid is a fan-shaped muscle. The geniohyoid pulls the hyoid bone forward and upward when it contracts, similarly to the mylohyoid muscle (Buchaillard et al., 2009; Shaw & Martino, 2013), leading to a slight elevation of the tongue (Epstein et al., 2002).

3.2.5 Muscles in the midsagittal ultrasound image

Muscles are usually hypoechoic, showing up as relatively darker pixels in ultrasound images (Carra et al., 2014; Reimers et al., 1993). For example, the genioglossus, which constitutes a large part of the inside of the tongue, shows up as the dark area marked with “A” in Figure 3.3. The tongue surface, by contrast, shows up as the bright curve marked with “B”. (In Figure 3.3 and all the images in the present study, the front of the mouth is to the right of the image.) The mylohyoid and geniohyoid are also visible as dark patches in the image (marked with “C”). Above and slightly to the left of the mylohyoid and geniohyoid, a moderately bright area close to the tongue root is visible (“D”). This area reflects the tongue fat, which is found predominantly in the tongue root (Kim et al., 2014; S. H. Wang et al., 2020). Above the mylohyoid and geniohyoid, slightly to the right, another bright area is located (“E”). It represents the tendon of the genioglossus (Wrench & Balch-Tomes, 2022; Wrench & Beck, 2022). Above the tendon is yet another bright area (i.e., “F”), which appears in ultrasound images when there is an air

pocket underneath the tongue tip.

A typical midsagittal ultrasound image usually contains two dark black big shadows at both sides of the image. The one located near the front of the mouth ("G") is called the mandible shadow, as it is a shadow created by the jaw (mandible). The dark area next to the tongue back is known as the hyoid shadow ("H"), which is created by the hyoid bone ("I").

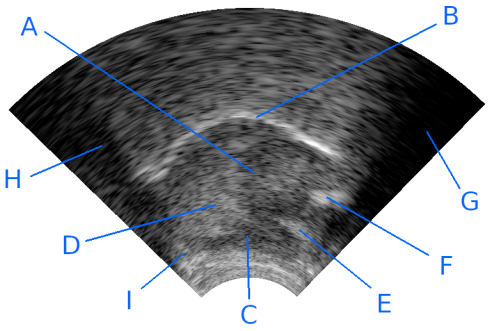


Figure 3.3: An example of an ultrasound image. A: genioglossus, B: tongue surface contour, C: mylohyoid and geniohyoid, D: Tongue fat, E: Tendon of genioglossus, F: Air pocket under the tongue tip, G: mandible shadow, H: hyoid shadow, I: hyoid bone.

3.2.6 Basic physics of ultrasound imaging

Ultrasound imaging is widely used in medical applications for creating images of internal organs. A sound emitting device known as the transducer sends out high frequency sound waves that cannot be perceived by the human ear. The transducer not only emits, but also receives ultrasound waves. The device operates with pulses, first emitting sound waves and then receiving their reflections. For the tongue recording, the transducer is usually placed under the chin.

Ultrasound pulses, emitted from the transducer, travel through the tongue and reflect at borders of substances with different acoustic impedances. The greater the differences in acoustic impedances are, the more of the ultrasound pulse is reflected back. Acoustic impedances are similar among different kinds of tissues, bones have greater acoustic impedances, and air has substantially reduced acoustic

impedances. Because of these differences in acoustic impedances, more than 99% of ultrasound pulses reflect at the border between the tongue and the air above the tongue. At the border of human tissue and bone, about 59% of ultrasound pulses reflect. Between different kinds of human tissues, very little of ultrasound pulses is reflected back.

Areas above air pockets and areas above bones show up as black shadows in ultrasound images. As the tongue tip often has an air pocket underneath it, it is often shadowed and invisible. Furthermore, part of the tongue root tends to be shadowed by the hyoid bone.

Ultrasound pulses are also attenuated by scattering when pulses encounter small targets or rough surfaces. Scattered ultrasound pulses appear as flickering white dots all over an ultrasound image. These dots are mostly at random locations in the image. However, certain constellations of irregularities in the tongue may give rise to artefacts. For example, an ultrasound pulse can reflect at one border and then, instead of being reflected straight back, it might vear off in a different direction, and then bounce straight back and then return to the transducer. The time required to travel back is now longer, which leads to the transducer to impute a location that is farther away than the actual border where the pulse was first reflected. In this case, the transducer mistakenly assumes that the received pulse travelled along a straight path and back again. As a consequence, a bright pixel is depicted in a wrong position. Such an artefact is known as a mirror-image artefact.

Another example of an image artefact is the reverberation artefact. This artefact can occur when relatively strong echos are created by a large smooth surface. The ultrasound pulse, reflected at such a surface, comes back to the transducer, where it can be reflected back again, resulting in a second wave that travels along the same path as the initial pulse. As such a second wave will arrive back at the transducer at a later point in time, it will be interpreted as being located at twice the actual distance. These kinds of artefacts lead to bright pixels far above the tongue surface. Without these artefacts, the area above the tongue border would be completely

black.

3.3 Modelling pixel brightness with GAMs

The Generalized Additive Model (GAM) is a flexible statistical tool for modeling a response variable as a non-linear function of one or more predictor variables. In the following example, the response is modeled with a general intercept β_0 and a smooth with two covariates, x_1 and x_2. $\hat{y} = \beta_0 + f(x_1, x_2) + \varepsilon$. When the smooth function f has a single argument, a wiggly curve is predicted. When it has two arguments, a wiggly surface is predicted. When there are more than two predictors, the result is a wiggly hypersurface. Smooths with longitude and latitude as predictors are widely used in biology (e.g., the density of sole eggs of the coast of Devon and Cornwall (Wood, 2017)) and dialectometry (Wieling et al., 2011). Applications are also found in linguistics (Baayen et al., 2010; Nieder, 2023) and psychology (Baayen et al., 2017; van Rij et al., 2019).

As ultrasound images are fully described by rectangular matrices with pixel brightness values, two-dimensional spline smooths can be applied to these matrices with as predictors the x and y coordinates, and pixel brightness as the response variable. In this study, we make use of tensor product splines, as these splines are ideal for models that separate out main effects and interactions.

Modeling with GAMs can be understood as an extension of the line of research using splines on the tongue contour (Davidson, 2006), and also as part of a line of research attempting to include as much information as possible from an ultrasound image (Palo, 2019, 2020, 2021).

In the following sections, we will show how GAMs can be used for analysing single ultrasound frames (Section 3.3.1), for comparing multiple ultrasound images (Section 3.3.2), and for predicting ultrasound images with an additional covariate such as time (Section 3.3.3). How speaker differences can be handled is discussed in Section 3.3.4.

3.3.1 Representing a single ultrasound image

A single frame of an ultrasound image is usually displayed in a fan-shape fashion as shown in Figure 3.4. This fan-shape picture, however, is the result of a geometric transformation with interpolation. Ultrasound pulses are emitted from the transducer and travel along their own "corridors" or ultrasound beams. Since these ultrasound beams spread from the transducer in a fan-shape fashion, ultrasound beams are more sparse as they get farther from the transducer and hence more interpolation is needed to reconstruct a fan-shaped picture.

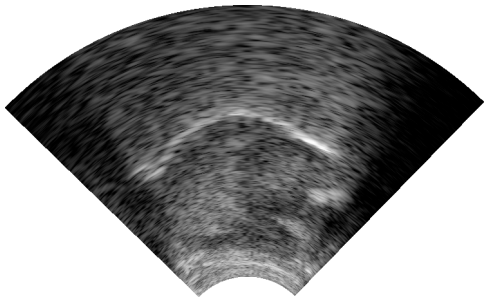
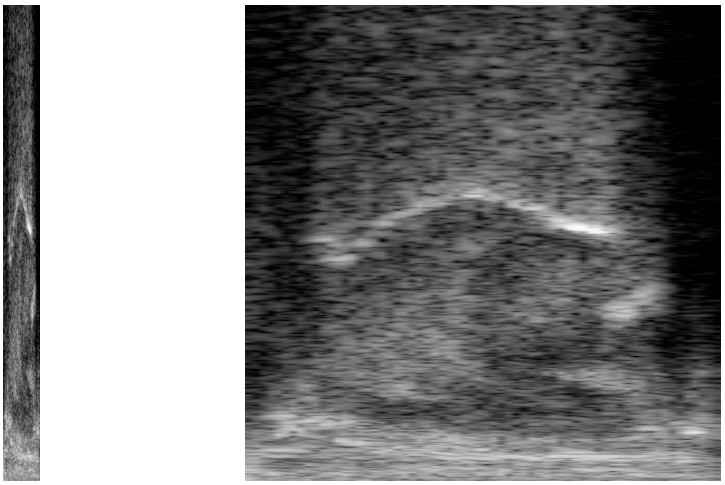


Figure 3.4: An example of an fan-shaped ultrasound image of German /a:/ in *ihr z**a**hlt*.

The data that are collected by the transducer for a given image is brought together in a single vector of unsigned integers that represent pixel brightness. This vector can be split into the sub-vectors representing the ultrasound scan lines, and these vectors can then be stored as the column vectors of a matrix representing the raw ultrasound image. An example of such an image is presented in Figure 3.5a. Figure 3.5b is the same ultrasound image, but stretched horizontally for better visibility.

The raw ultrasound image is a simple square grayscale image. In grayscale images, pixel brightness is represented by unsigned integers ranging from 0 to 255. Therefore, pixel brightness can be modelled with a tensor product smooth using the x- and y-coordinates as predictors.

The simplest GAM model for a single ultrasound image (using the notation of the **mgcv** package (Wood, 2017) for R (R Core Team, 2022)) specifies a tensor



(a) Raw ultrasound image. (b) Raw ultrasound image resized to a square.

Figure 3.5: Examples of a raw ultrasound image corresponding to Figure 3.4 (left) and the same figure but stretched horizontally for better visibility (right).

product smooth.

```
brightness ~ te(x,y)
```

In this specification, an intercept is automatically included during model fitting. A tensor product smooth can also be split into main effects and an interaction term, modeled with the `ti` directive, as follows:

```
brightness ~ s(x) + s(y) + ti(x,y)
```

Here, `s(x)` represents a thin plate regression spline smooth for x.

The wiggliness of spline curves and tensor product smooths is controlled by the number of basis functions used to construct the smooths. The number of basis functions is specified in `mgcv` by an optional argument `k`. Greater values of `k` enable more precise modeling of nonlinear trends. The default of k is not theoretically motivated, and the analyst has to make sure that k is set to a value that is large enough for adequate modeling. For a given k, the GAM algorithm penalizes the coefficients of the basis functions in order to optimally balance oversmoothing and undersmoothing.

For ultrasound images, which are characterized by sudden changes in brightness, we use 20 basis functions for both coordinates, and set to 20 for both dimensions (`k=c(20,20)`):

```
brightness ~ te(x, y, k=c(20,20))
```

When the raw ultrasound image shown in Figure 3.5a is fitted with this model, the fitted surface presented in Figure 3.6 is obtained. In this figure, the front of the mouth is to the right, following standard conventions for raw ultrasound images. Warmer colors represent higher pixel brightness values. The axes cover the same ranges of values as in Figure 3.5a, but the axis are rescaled for better interpretability.

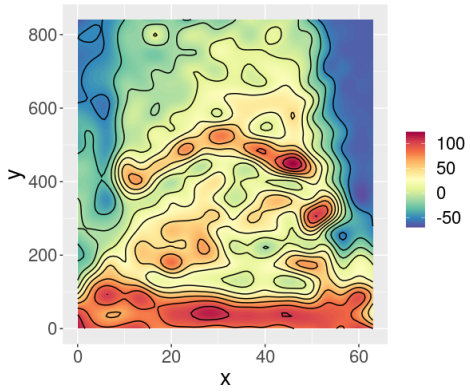


Figure 3.6: The fitted surface for Figure 3.5a.

Figure 3.6 is the fitted surface for a raw ultrasound image. It has to be transformed to the usual fan-shape with interpolation in order to correctly represent the tongue shape. This can be achieved with the same geometric transformation and interpolation method that was used above to transform a raw ultrasound image into its corresponding fan-shape image:

Figure 3.7 corresponds to 3.4. The bright wedge-shape curve in the center of the figure represents the tongue contour. The yellow-colored areas above the tongue surface are artefacts, that arise for the reasons discussed in the preceding section.

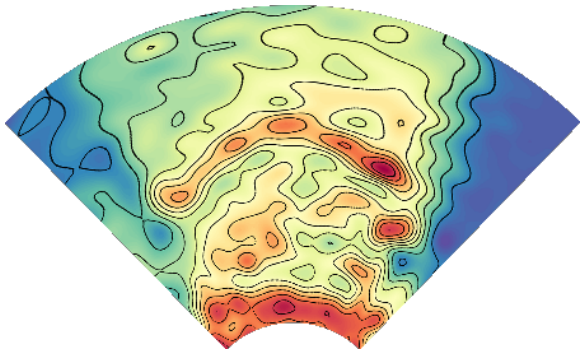


Figure 3.7: The fitted surface for Figure 3.5a, transformed to the fan-shape.

Below the tongue surface, there are some areas that are unlikely to be artifacts. First, below the tongue tip, is a small red area that highlights a pocket of air below the tongue tip. The bright area at the bottom of the GAM contour plot represents the fat and skin below the mylohyoid and the geniohyoid. Between the bright area in the bottom and the tongue contour is another moderately bright and relatively large area slightly to the left in the contour plot. This area represents tongue fat that is located more towards the tongue root.

In Figure 3.7, there are two big shadows to the left and the right of the image. The one to the left of the image, the hyoid shadow, is created by the hyoid bone located in the tongue root. The shadow to the right of the image is created by the mandible (jaw) bone.

The visualization of the tensor product smooth in Figure 3.7 clearly reveals where the tongue surface is located. Furthermore, the tongue surface is clearly separated from the pocket of air below the tongue tip. Unlike a widely used method for tracking the tongue surface (EdgeTrak: M. Li et al., 2005) which requires pre-processing of images by the analyst, the GAM provides a clear regression surface that does not require prior selection of regions of interest.

In practice, the analyst will be interested in comparisons between ultrasound images, and specifically in differences between locations of the tongue surfaces as a function of experimental treatments. The next subsection explains how GAMs can be used to address these questions.

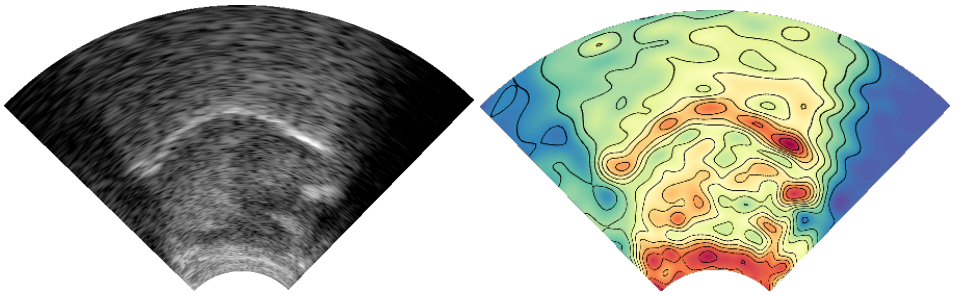
3.3.2 Comparing two ultrasound images

Before introducing the construction of spline surfaces that directly represent differences, we first need to introduce how to model interactions of x- and y-coordinates with a factorial predictor. Consider, for instance, two images of the tongue at the center of a vowel, as realized in German verbs, that differ only with respect to their morphological structure: the third person singular and the first/third plural (*zahlt*, 'he/she/it pays' vs. *zahlen*, 'we/they pay'). We would like the GAM to fit two brightness surfaces, one for the vowel [aː] of *zahlt* and one for that of *zahlen*. The analyst can request a separate regression surface for each of the levels of the factorial predictor `morphology`, using the `by` directive in the call to `gam`:

```
brightness ~ morphology +
             te(x, y, k = c(20, 20), by = morphology)
```

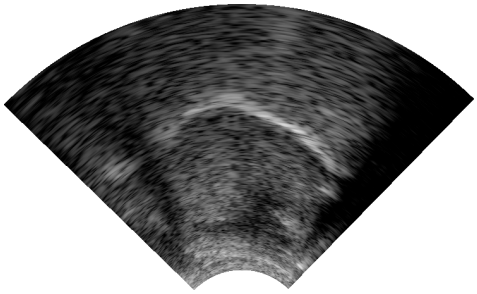
Note that the factorial predictor `morphology` is included also as a main effect in the above formula. Its inclusion ensures that a possible difference in the means of the two surfaces is taken into account. In the case of ultrasound images, differences in means represent overall shifts of brightness across ultrasound images. For example, the ultrasound image for *zahlt* (Figure 3.8a) appears to be slightly brighter overall, compared to that for *zahlen* (Figure 3.8c), possibly due to changes within the tongue giving rise to different scattering of ultrasound beams. As indicated in Table 3.1, `morphology` has an estimated coefficient of 13.959, indicating that the ultrasound image for *zahlt* is brighter by about 14 brightness values (which range between 0 and 255) on average than that for *zahlen*.

In Figure 3.8, the left two panels present the ultrasound images for *zahlt* (top) and *zahlen* (bottom). The smooths to the right of these grayscale images present the fitted surfaces that are estimated by the GAM. The fitted surfaces, by their nature, include the intercept and the adjustments to the intercept that come with the main effect of `morphology`. The fitted surfaces are therefore directly comparable with the raw ultrasound images to their left.

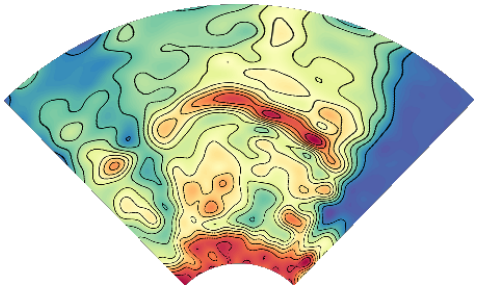


(a) Input, *zahlt*

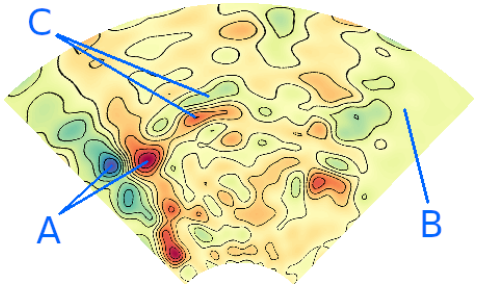
(b) Prediction, *zahlt*



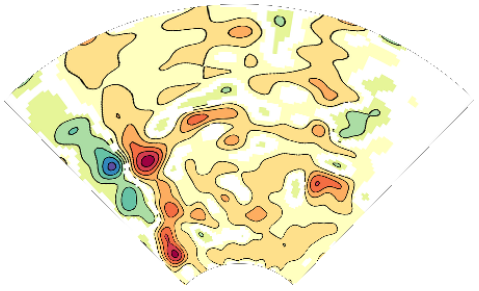
(c) Input, *zahlen*



(d) Prediction, *zahlen*



(e) Predicted differences



(f) Predicted differences with insignificant differences being blank

Figure 3.8: Input (left) and predicted (right) ultrasound images for the stem vowel (i.e., [aː]) of *zahlt* [tsaːlt] (top) and *zahlen* [tsaːlən] (second row). The figure in the third row represents the differences between the predicted images of *zahlt* and *zahlen*. The figure in the bottom row is the predicted differences with insignificant differences between the two conditions being blank. The areas that are marked by 'A' show that there are differences in positions of the hyoid shadow between *zahlen* and *zahlt*. 'B' indicates that there is no difference between the conditions. 'C' suggests that the tongue body positions are slightly different with the tongue body for *zahlen* being slightly higher.

Table 3.1: The summary of the model implementing two surfaces for the two morphological conditions.

A. Parametric terms	Estimate	Std.Error	t-value	p-value
Intercept	61.296	0.109	563.080	<0.001
morphology=zahlt	13.959	0.154	90.736	<0.001
B. Smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
te(x, y):morphology=zahlen	386.166	397.845	433.559	<0.001
te(x, y):morphology=zahlt	382.264	396.998	577.649	<0.001

The red curved areas in the middle of the fitted surfaces, seen in Figures 3.8b and 3.8d, correspond to the bright (white) curves seen in their corresponding raw ultrasound images (i.e., Figures 3.8a and 3.8c), which represent the tongue surface. The estimated tongue surfaces look wider in the fitted ultrasound images than in the observed ultrasound images. One reason is that the tongue surface in the fitted surfaces reflect uncertainty about the location of the tongue surface. Another reason is that GAMs work with smooths that need to be differentiable and hence cannot deal with abrupt discontinuities, such as a transition from completely black to completely white.

In general, summaries of GAM models provide the analyst with information about the significance of partial effects. For the present model, all the terms are significant, including the parametric term `morphology` and two surfaces created as a function of x- and y-coordinates for each level of `morphology` (Table 3.1). The tests for the two smooth surfaces are not very revealing, because there are obvious differences in pixel brightness all over the image, and many of such differences are devoid of theoretical interest. What is more useful to the analyst is a surface that is informative about where two images differ, and whether the observed differences are present in areas that are of interest, such as the tongue surface.

For comparing differences between two levels of a categorical variable, there are two possibilities, regarding how to treat overall shifts in brightness between conditions. For the current model, the parametric term is significant (Table 3.1)

and indicates that there is a significant overall shift in brightness between the two images being compared. Such an overall shift in brightness between conditions may be systematic and informative, but it may also be basically random and of no theoretical interest. If such an overall shift in brightness is not of theoretical interest and assumed to be random, then the response variable (i.e., brightness values of pixels) can be centered and scaled before fitting a GAM model. This procedure will eliminate an mean difference between conditions (e.g., `morphology`). Since the mean difference is taken away beforehand, no parametric term is necessary in this case to code for different conditions.

While no parametric term is necessary, the analyst still has to set up the model in such away that it captures two input ultrasound images properly while bringing out the differences between them. For this goal, the analysis first needs to set up a reference surface, for instance, the surface for *zahlen*. This surface will be wrong for the other level, *zahlt*. Therefore, the GAM model will need to add a surface that changes the reference surface for *zahlen* into the correct surface for *zahlt*. This 'correcting surface' is the difference surface. To this end, the categorical variable of interest, here `morphology`, must be explicitly coded as a numeric binary variable `morphology_num`, with 0 corresponding to the reference level of `morphology` (i.e., *zahlen*) and 1 to the treatment level (i.e., *zahlt*). Denoting centered brightness by `c.brightness`, a GAM model can now be estimated as follows:

```
c.brightness ~ te(x, y, k = c(20, 20)) +
               te(x, y, k = c(20, 20), by = morphology_num)
```

The summary of this model is presented in Table 3.2.

The first tensor product smooth (i.e., `te(x, y, k = c(20, 20))`) fits a regression surface for the reference level of `morphology_num` (i.e., `zahlen`). The second tensor smooth (i.e., `te(x, y, k = c(20, 20), by = morphology_num)`) fits the difference surface. When this difference surface is added to the reference surface, the surface for `zahlt` is obtained.

Table 3.2: The summary of the model implementing a difference surface for the morphological conditions.

A. Parametric terms	Estimate	Std.Error	t-value	p-value
Intercept	0.008	0.002	4.746	<0.001
B. Smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
te(x, y)	387.501	397.525	489.350	<0.001
te(x, y):morphology_num=1	379.062	396.096	64.221	<0.001

Note that the first tensor product smooth (i.e., the reference surface) does not have the `by` option, while the second tensor product smooth (i.e., the difference surface) does. Because the reference surface does not have the `by` option, it is always "on", regardless of the levels of the `morphology_num`. In contrast, the difference surface is specified with a `by`-directive. When `by` points to 0, the smooth is pre-multiplied with 0, and thus effectively cancelled. As a consequence, a 'zero-difference smooth' is added to the smooth for the data points of the reference level. When `by` points to 1, the second (difference) smooth is multiplied with 1, and hence retained. Thus, when the treatment level (i.e., `zahlt`) is estimated, the first and second tensor terms are both used. The regression surface for `zahlt` is obtained by taking the regression surface for the reference level (`zahlen`) and adding to this the difference surface for the treatment level `zahlt`.

This model does not need to include the parametric term for `morphology` or `morphology_num`, since the dependent variable is centered and scaled prior to fitting the model (i.e., `c.brightness`), eliminating mean differences between conditions effectively. This pre-processing is based on the assumption that any mean difference between conditions should be at random and not of any theoretical interest.

In contrast, if one wishes not to have such an a priori assumption about overall differences in brightness, the parametric term can be brought back to the formula without centering and scaling the response variable (i.e., `brightness`):

```
brightness ~ morphology_num +
```

```
te(x, y, k = c(20, 20)) +
te(x, y, k = c(20, 20), by = morphology_num)
```

This model with the parametric term provides a significantly better model fit than the model without the parametric term for the current dataset (ΔAIC $= 8782.41$; ΔML $= 4441.324$, $\chi^2(1) = 4441.324$, $p < 0.001$). This difference in model fit is likely due to the fact that, in the model without the parametric term, the overall shift in brightness had to be absorbed by the difference smooth by itself. This smooth does not include a 'horizontal' basis function (which for identifiability reasons is merged into the intercept). As a consequence, more basis functions are required, leading to a higher AIC.[2](Baayen & Linke, 2020). On the other hand, the overall shift in brightness was taken into account by the parametric term in the model with the parametric term. In other words, the task of estimating differences between surfaces of different conditions is more difficult for the model without the parametric term than for the model with the parametric term.

The significance of the difference surface in the summary table (Table 3.3) clarifies that the two regression surfaces differ significantly. The difference surface is visualized in Figure 3.8e. The difference surface is colored in such a way that warmer colors represent brighter pixels in the *zahlt* condition than in the *zahlen* condition.

In the center left of the difference surface there are a deep red area and a dark blue area next to each other (marked "A"). They together indicate that the position of the hyoid shadow is different between the two conditions. In the case of *zahlen*, the hyoid shadow is located more towards the center.

It is noteworthy that this movement of the hyoid shadow is not due to rotation of the transducer. This can be deduced from the mandible shadow, which hardly

[2]The first basis function is usually not included in a GAM model due to the identifiability problem. If a smooth term also has its own intercept (represented by the first basis function) in addition to the "grand" intercept, then the coefficient for this intercept can be increased by a, and the grand intercept can be decreased by a, without changing model predictions. In this case, unfortunately, there will be infinitely many models that are identical in model performance, and it will not be possible to determine which model to choose.

Table 3.3: The summary of the model implementing a difference surface with the parametric term for the morphological conditions.

A. Parametric terms	Estimate	Std.Error	t-value	p-value
Intercept	61.660	0.107	577.555	<0.001
morphology_num=1	15.118	0.151	100.266	<0.001
B. Smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
te(x, y)	386.072	397.229	433.889	<0.001
te(x, y):morphology_num=1	378.688	395.946	69.420	<0.001

changes between the two conditions: There are no systematic differences in pixel brightness at the right-hand side of the difference surface (marked "B"). If the movement of the hyoid shadow were mainly due to rotation of the transducer, the mandible shadow should also have moved to the same degree as the hyoid shadow, contrary to fact. Since the mandible shadow is fixed and the hyoid shadow is fronted, the areas "A" and "B" together indicate contraction of the mylohyoid, the geniohyoid, and possibly also the posterior genioglossus. By contracting, these muscles bring the hyoid bone forward and upward, squeezing the bottom of the tongue, and resulting in raising of the tongue body.

The green and red areas highlighted by "C" represent the differences in tongue surface positions of the posterior part of the tongue. The red area shows where the ultrasound image is brighter for *zahlt*. The green area above it indicates where the ultrasound image is brighter for *zahlen*. Thus, the difference surface shows that the position of the posterior part of the tongue surface is slightly pushed up for *zahlen*, compared to *zahlt*. Furthermore, the rightward shift of the hyoid shadow indicates that the tongue root is more fronted, so that the tongue is also more bulged for *zahlen*, compared to *zahlt*.

The difference surface simply shows what differences exist where. Since this difference surface is a predicted regression surface, it comes with its own confidence intervals. When the confidence interval of a difference surface contains 0, then there is no significant difference between conditions. However, the fitted val-

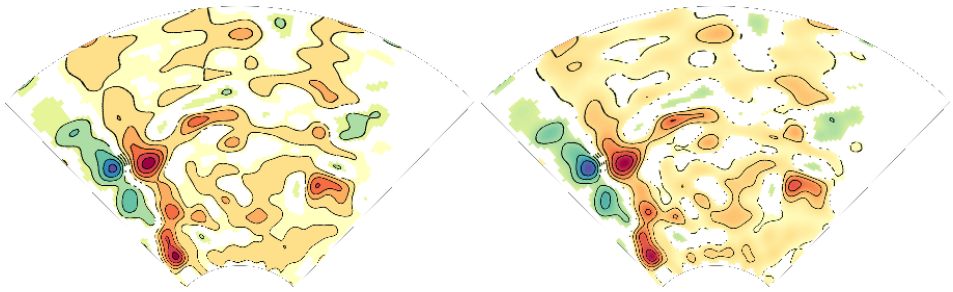
ues and the confidence intervals of the difference surface should be taken with caution in this setting, due to the presence of the parametric term for `morphology` (i.e., `morphology_num`). Such a parametric term indicates that one surface is higher or lower than the other *on average*. However, for regions where there is no difference between the two conditions, as is the case for the mandible shadow, the difference smooth is forced to compensate for the mean difference, resulting in negative values around -15 for the difference surface in the current dataset. In this way, the predicted values for the regression surfaces for both `zahlt` and `zahlen` will be zero.

The drawback of all this is that the difference surface will be significantly below zero for the mandible shadow, even though the two surfaces have zero brightness and are not different at all. To avoid this problem, one option is to add the parametric term to the difference surface, and then consider whether the confidence interval of this shifted surface contains zero. Figure 3.8f was produced following this procedure. Areas for which the confidence interval includes zero are represented in white.

A disadvantage of this procedure is that addition of the parametric term may introduce a significant 'main effect' difference for other areas even when brightness values for the two surfaces are very similar. As a consequence, interpretation of the difference surface becomes more complex.

An alternative is to focus on the fitted surfaces and their 95% confidence intervals. In general, non-overlapping of confidence intervals is a strong indication of a significance difference, while overlapping does not necessarily suggest insignificance (Cornell Statistical Consulting Unit, 2020). We therefore implemented a fairly conservative assessment of the difference surface, accepting regions as significantly different only when the 95% confidence intervals of the two predicted regression surfaces do not overlap. In plots visualizing the difference surface in the following sections, all areas where 95% confidence intervals overlap are displayed in white.

Figure 3.9 presents the two difference surfaces, so that the results of the two methods can be compared. The left panel presents the same difference surface as shown in Figure 3.8f, the only difference being that now differences were evaluated by the confidence intervals at a significance level of 0.001, instead of 0.05. The right panel presents the difference surface with whitening wherever the confidence intervals of the two fitted surfaces overlap ($\alpha = 0.05$). Figure 3.9 shows that the two methods provide very similar results.



(a) By the CIs of the difference surface plus the intercept.

(b) By overlaps of CIs of the two surface under comparison.

Figure 3.9: Visualization of the estimated difference surface with non-significant areas being blank, evaluated by the intersection of the CIs of the shifted difference surface with 0 at the $\alpha = 0.001$ level (left) and by overlaps of the CIs of the two surfaces under comparison at the $\alpha = 0.05$ level (right).

As was the case for the evaluation method by the confidence intervals of the difference surface, this method focusing on overlaps of fitted surfaces also show that the differences due to the movement of the hyoid bone are well supported, as well as the small difference in tongue body position. In addition, similarly to the method of the difference surface (Figure 3.9a), the evaluation based on overlaps of confidence intervals (Figure 3.9b) also shows several orange areas above the tongue surface, for which the method reports medium differences in brightness. These differences are not of theoretical interest as they most likely reflect the differences in the extent to which ultrasound beams were scattered.

Ideally, one would want to compare surfaces from the same model. However, due to the large number of data points in ultrasound images, in the case study we

report below, we fitted separate models and evaluated similarities and differences by comparing the absence of overlap between confidence intervals as a guide to areas of interest.

3.3.3 Including covariates as predictors

In the previous subsection, we included a categorical variable (i.e., `morphology`), using the `by` directive. A continuous variable such as `time` can also be included, for instance, using a three-dimensional tensor product smooth.

```
brightness ~ te(x, y, time)
```

When multiple consecutive frames are analyzed, bright pixels at one frame tend to be still bright at the next frame. This autocorrelation should be taken into consideration when time course data is analyzed. It can be achieved by including an AR1 process for the errors (Wood, 2017), according to which the current error at time t is a proportion of the error at time $t-1$ plus gaussian noise.

As an example, ultrasound images of the stem vowel of *zahlt* ([tsaːlt]) were fitted with the three-way tensor product smooth and AR1 process with $\rho \approx 0.372$. For visualization, five frames were selected at the quartiles, i.e., 0%, 25%, 50%, 75%, and 100% into the vowel.

Figure 3.10 shows the resulting five fitted surfaces together with the corresponding ultrasound images. Going from top to bottom, the raising of the tongue tip is clearly visible, as well as the flattening of the tongue surface, suggesting contraction of the medial genioglossus. The raising of the tongue tip reflects anticipatory coarticulation with the upcoming alveolar consonants ([-lt]), and is likely due to be supported by the superior longitudinal muscle.

In addition, it is worth noting that the hyoid shadow is fully visible in the top panels, but has shifted somewhat to the left by the time the center of the vowel is reached. The initial elevation and advancement of the hyoid bone, that result in a rightward shifted hyoid shadow, are likely due to the contraction of the mylohyoid

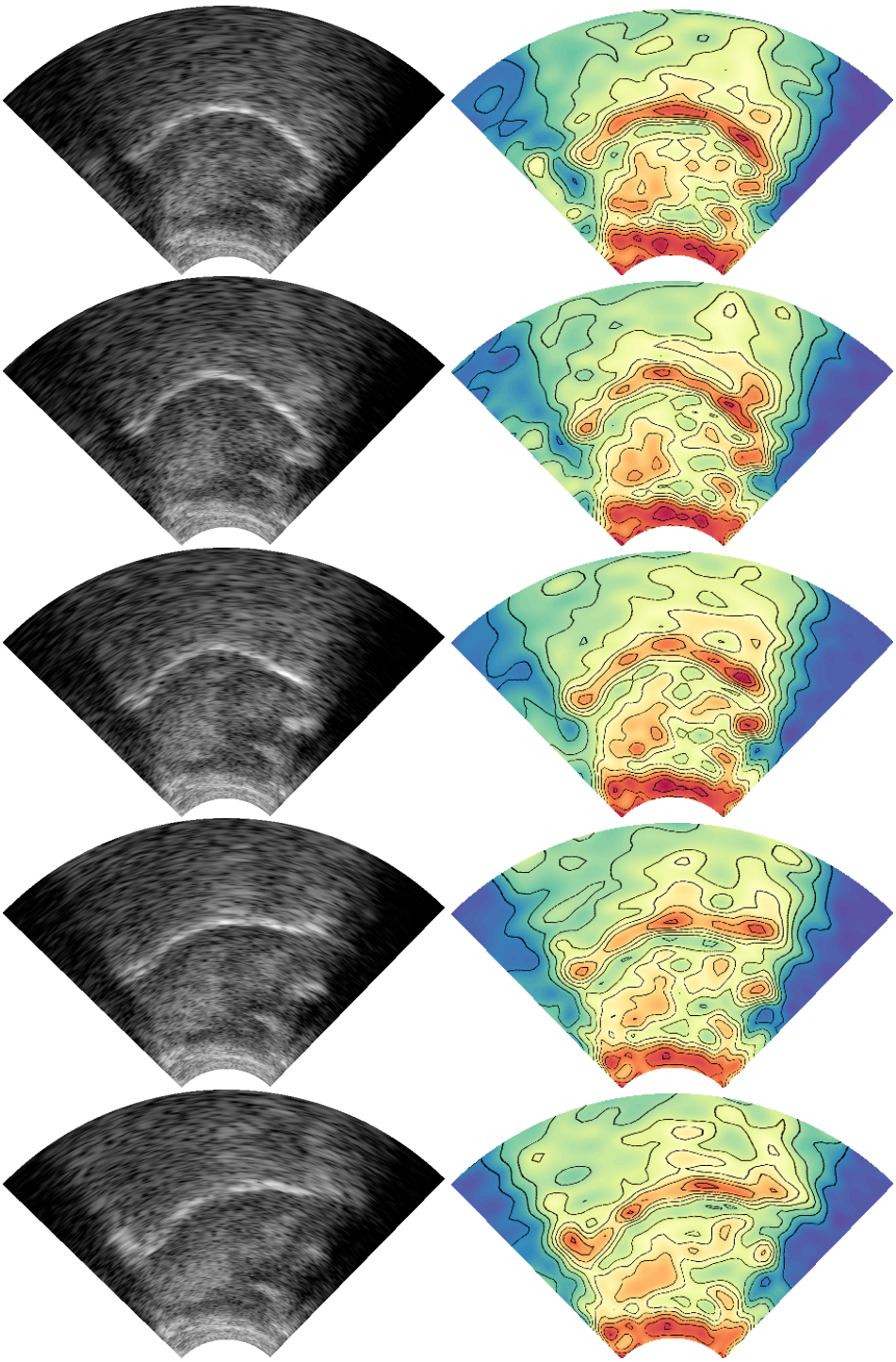


Figure 3.10: Development of the tongue shape during [aː] in *zahlt* [tsaːlt] from the onset (top) to the offset (bottom).

and the geniohyoid muscles (Epstein et al., 2002). In the top two frames of Figure 3.10, the contraction of these two muscles results in elevation of the floor of the mouth and contribute to the bulging of the tongue body. Thus, the upper panels present the initial configuration of the tongue for the vowel, although it cannot be ruled out that there is some co-articulation with the preceding segments.

Another covariate can be included in the same way as `time` was included above. For example, it has been known that duration and frequency can both affect movements of the tongue (Dinkin, 2008; Kelso et al., 1985; Kuehn & Moll, 1976; Lin et al., 2011; Tomaschek, Arnold, et al., 2018; Tomaschek, Tucker, et al., 2018; Tomaschek et al., 2021). In order to include `duration` and `frequency` as covariates, one might want to simply set up separate tensor product smooths for each of the covariates as below:

```
brightness ~ te(x,y, duration) + te(x,y, freq)
```

However, this specification is not correct, because the two terms are competing in part for the same variance. What we need is a completely decompositional model, in which main effects and interactions are carefully distinguished. Interactions of two or more numeric variables have to be fitted with the `ti` directive, rather than the `te` directive. The `ti` terms are appropriate for functional ANOVA decomposition and provide interaction (hyper)surfaces from which the main effects have been excluded. In the following model, the main effects are specified first, followed by all pairwise interactions and the two three-way interactions of interest between the x- and y-coordinates and the two additional covariates.

```
brightness ~ s(x) + s(y) + s(dur) + s(freq) +
             ti(x,y) + ti(x,dur) + ti(x,freq) +
             ti(y,dur) + ti(y,freq) +
             ti(x,y,dur) + ti(x,y,freq)
```

A main effect term such as `s(freq)` specifies how pixel brightness varies with frequency, irrespective of position in the image. An interaction term such as `ti(x,`

`freq`) allows the effect of frequency to vary in the direction of the x-axis, regardless of the y-axis position. A three-way interaction such as `ti(x,y,freq)` captures changes in brightness by frequency in the x-y plane that are not captured by the main effects and lower order interactions. Thanks to this ANOVA decomposition, it is possible to include multiple covariates and their interactions with the x and y coordinates.

In this subsection, we have sketched how convariates such as `time` can be brought into a GAM model. The next subject shows how differences between speakers can be taken into account.

3.3.4 Speaker as random effect

GAMs can include random-effect factors. However, for ultrasound data, by-speaker random intercepts only allow for differences in overall pixel brightness. But what we are interested in is speaker-specific modulations of the regression surface. To capture different degrees and patterns of wiggliness for different speakers, we can request the GAM algorithm to allow each level of speaker to have its own wiggly curve. As there are many different speakers, it may be desirable to treat speaker as a random effect factor. In the context of GAMs, this means we assume that the amount of wiggliness in the speaker-specific partial effects is roughly the same. Wiggly random effects are requested by setting the `bs` option to the value ''`fs`'' (i.e., factor smooth):

```
brightness ~ te(x, y) + te(x, y, speaker, bs = "fs")
```

In this model specification, the first tensor product corresponds to the predicted surface (i.e., predicted ultrasound image) that is common to all speakers.

For demonstration, we focused on the production of the stem vowel [aː] of *ihr zahlt* [ɪɐ tsaːlt] by two speakers. Different speakers usually have different sizes of the oral cavity and the tongue. To normalize these differences, we took the following four points as reference point for cropping the images prior to analysis: (a)

the border of the tongue contour line and the hyoid shadow, (b) the border of the tongue contour line and the mandible shadow, (c) the highest tongue point during articulating [k], (d) the border of the skin and the mylo/genio-hyoid muscles. We considered the articulation of [k] to estimate the position of the palate, because [k] constantly showed higher tongue constriction points than [t]. This cropping procedure normalizes different sizes of the oral cavity across speakers, with the left and right edges of the image corresponding to the hyoid shadow and the mandible shadow respectively, and with the top and bottom of the image corresponding to the palate and the mylo-/genio-hyoid respectively. These cropping points are illustrated in Figure 3.11.

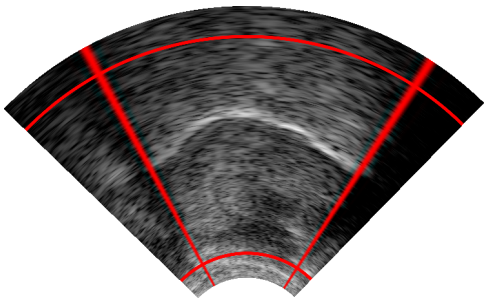
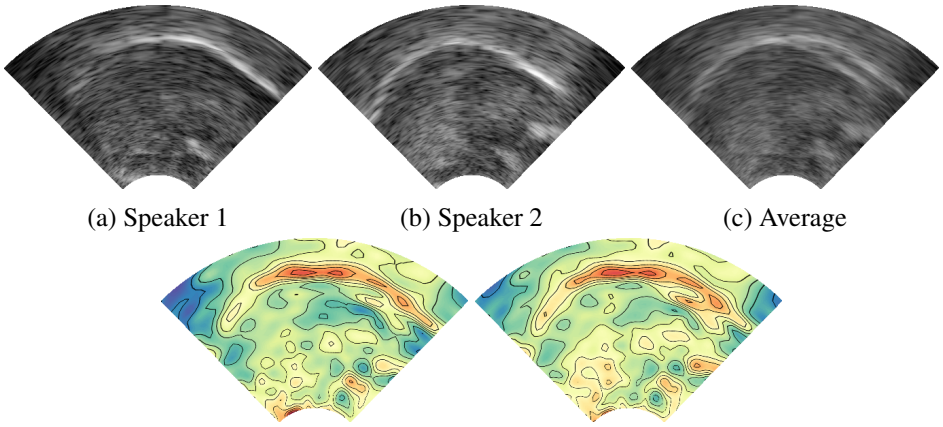


Figure 3.11: Example of the selection of the area to be included in analyses with multiple speakers, which requires normalization for different sizes of the oral cavity.

The first two figures in Figure 3.12 are the cropped ultrasound images of the two speakers. The tongue position of Speaker 1 is slightly more fronted than that of Speaker 2. In addition, the tongue tip is also higher for Speaker 1. These two figures were provided as input to a GAMM with the model specification presented above.

The partial effect of the main tensor product is shown in Figure 3.12d. This main tensor product captures the average of the ultrasound surfaces of the two speakers. When the average of the raw ultrasound images is computed (see the third panel on the top row), we obtain a surface that is very similar to the partial effect of the main tensor product. The lower right panel presents the predicted sur-



(a) Speaker 1 (b) Speaker 2 (c) Average

(d) By a factor smooth (e) By averaging images

Figure 3.12: Ultrasound images at the middle of [aː] in *ihr zahlt* from two speakers (a,b), the averaged image between the two (c), and predicted ultrasound images by a factor smooth (d) and by averaging images in prior to fitting a GAM.

face of a GAM fitted to the average of the two ultrasound images. As expected, the two predicted surfaces are very similar. Because main effect tensor smooths provide central tendencies across different speakers, areas that are consistently bright across all speakers will show up with high predicted brightness values, whereas areas where speakers are highly variable, brightness values will be reduced. In Figure 3.12d, the position of the tongue middle/body shows up with shades of red, as for both speakers, this part of the tongue is in a very similar position. However, the position of the back/root of the tongue differs for the two speakers, and accordingly this part of the tongue is represented by lighter colors. In other words, when more than one speaker is included, degrees of predicted brightness can also indicate variability among speakers. In addition, variability among speakers can also show up as a larger "area" of the tongue surface position. Although we focused on speaker differences so far, this discussion about variability among speakers can also be applied to differences across frames and items.

Our discussion thus far illustrates how speaker variability can in principle be integrated into a GAM model. Unfortunately, estimating by-speaker random wiggly curves becomes prohibitively computationally expensive for larger numbers

of speakers. To reduce computational load, we therefore averaged images across speakers, and did not include speaker as a separate random effect term in the models reported below.

3.4 Case study: Enhancement effects of frequency

In this section, we use GAMs to study possible effects of frequency on pronunciation. Frequency of occurrence has been investigated intensively in psycholinguistics. Oldfield and Wingfield (1965) found that word naming was slower for lower frequency words. Since then, many studies have reported effects of frequency across tasks, methodologies, languages, and different types of items (Baayen et al., 1997; Baayen et al., 2006; Baayen et al., 2002; Bertram et al., 2000; Forster & Chambers, 1973; Gahl, 2008; Gardner et al., 1987; Rubenstein et al., 1970; Scarborough et al., 1977; Schreuder & Baayen, 1997; Whaley, 1978; Wurm et al., 2006).

Higher frequency words have widely been reported to show more phonetic reduction, namely shorter duration, more centralized formant realizations, and less clear articulations (Aylett & Turk, 2004; A. Bell et al., 2009; A. Bell et al., 2002; Dinkin, 2008; Jurafsky et al., 2001; Munson & Solomon, 2004; Pluymaekers et al., 2005b). Furthermore, higher frequency words are more likely to be encountered and are more predictable in context. Higher predictability, in turn, goes in hand in hand with greater redundancy and a reduced information load (Aylett & Turk, 2004; Jurafsky et al., 2001). The amount of information carried by a word is defined as the negative logarithm (usually with base 2) of its probability of occurrence (Shannon, 1948). Words that carry little information are relatively redundant and this has been argued to underlie the extent to which such words undergo phonetic reduction.

While the reduction effect of frequency has repeatedly been replicated, frequency effects in the opposite direction have also been reported. Kuperman et al.

(2007) investigated the duration of interfixes in Dutch compounds and found that interfix duration was longer, rather than shorter, when the interfix was predicted better in the morphological paradigm the compound word belongs to and when the compound word that carries the interfix was a high frequency word. The possibility that a greater paradigmatic probability goes hand in hand with phonetic strengthening is supported by several subsequent studies (M. J. Bell et al., 2021; Cohen, 2014; Tomaschek, Tucker, et al., 2018; Tomaschek et al., 2021).

Tomaschek, Tucker, et al. (2018) investigated the stem vowel [aː] of German inflected verbs and found that frequency effects were modulated by the suffixes following the stem vowel. When the stem vowel was followed by the suffix *-t*, it was articulated with lower tongue trajectories as frequency of the word increased, indicating a phonetic enhancement effect of frequency. However, when the suffix *-en* followed the vowel, no clear effect of frequency was present.

In addition, Tomaschek, Tucker, et al. (2018) observed that the enhancement effect of frequency was non-linear. The greatest articulatory reduction was observed for words with average frequency. In contrast, both high frequency words and low frequency words were articulated with lower, more enhanced, tongue tip trajectories. In comparison to low frequency words, high frequency words tended to be articulated with earlier initiation of tongue raising towards the offset of the stem vowel, anticipating the upcoming alveolar suffix. In other words, low frequency words were articulated in such a way that clarity of the vowel was maximized. Middle frequency words were articulated the most smoothly. High frequency words realized both clarity and smoothness. Tomaschek, Tucker, et al. (2018) interpreted these results as reflecting articulatory optimization, made possible by more extensive motor experience. This interpretation was supported by a further study (Tomaschek, Arnold, et al., 2018), which reported that for high frequency words more complex tongue trajectories are articulated without a loss of speed.

In what follows, we report a replication study, using ultrasound instead of

EMA, of the frequency effects reported by Tomaschek, Tucker, et al. (2018), using an extended data set of German inflected verbs. The focus of this replication study is the difference in articulation between high and mid frequency words. Tomaschek, Tucker, et al. (2018) found that middle frequency words were articulated with shallower tongue trajectories, while high frequency words were articulated with lower tongue trajectories.

The materials in Tomaschek, Tucker, et al. (2018) contained 27 German verb types. These verb types had [aː] as the stem vowel and were inflected and combined with the third person plural pronoun *sie* and its corresponding suffix *-en*. Nine of these verbs were also combined with the second person plural pronoun *ihr* and its corresponding suffix *-t*, as illustrated below:

(13) sie zahlen. [ziː tsaːl(ə)n]
they pay.3PL
'They pay.'

(14) ihr zahlt. [ɪɐ tsaːlt]
you.PL pay.2SG
'You (plural) pay.'

Note that, in this study by Tomaschek, Tucker, et al. (2018), the suffix *-en* was always combined with the pronoun *sie*, and the suffix *-t* was always combined with the pronoun *ihr*. As a consequence, the possibility cannot be excluded that systematic differences in the articulation of [aː] found by Tomaschek, Tucker, et al. (2018) were confounded by the systematic differences in pronouns, namely the carryover coarticulation (Öhman, 1966; Repp & Mann, 1982; Song et al., 2013).

While the pronouns *sie* and *ihr* were always combined with the suffixes *-en* and *-t* respectively in Tomaschek, Tucker, et al. (2018), *sie* can also be used as the third person sinular pronoun by being combined with the suffix *-t*. In addition, the rime segments of *ihr* [iːɐ] can also occur as the rime of the first person plural pronoun *wir* [viːɐ], being combined with the suffix *-en*. Furthermore, the suffix *-t* can be singular or plural, depending on pronouns, and the suffix *-en* can be used as the

first person plural and the third person plural suffix both, as shown below:

(15) sie zahlt. [ziː tsaːlt]
he/she/it pay.3SG
'He/she/it pays.'

(16) wir zahlen. [vɪɐ tsaːl(ə)n]
we pay.1PL
'We pay.'

Table 3.4 shows all the possible suffixes for German present tense indicative inflected verbs.

Table 3.4: Inflectional exponents of the German verbs in the present tense

Person	Singular	Plural
1st	-e	-en
2nd	-st	-t
3rd	-t	-en

In order to control possible carryover coarticulation from the preceding pronoun, we adopted additional two inflectional variants, namely the pronoun *sie* with the suffix *-t* and the pronoun *wir* with the suffix *-en*, as summarized in Table 3.5 below:

Table 3.5: Combinations of pronouns and suffixes of interest with *sagen* [zaːg(ə)n] as an example.

	Suffix	
Pronoun	[-t]	[-(ə)n]
sie [ziː]	*sie sagt*	*sie sagen*
ihr/wir [(v)ɪɐ]	*ihr sagt*	*wir sagen*

3.4.1 Method

Participants

A total of 18 participants took part in the experiment. These participants were students at the university of Tübingen and received 10 Euro in compensation for their participation. All participants had normal hearing.

Materials

The dataset contained 395 target phrases, 52 of which belonged to the `sie-t` condition (e.g., *sie zahlt*), 152 to the `sie-n` condition (e.g., *sie zahlen*), 116 to the `ihr-t` condition (e.g., *ihr zahlt*), and 75 to the `wir-n` condition (e.g., *wir zahlen*). The verbs of these target phrases contained [a] or [aː] as the stem vowel. In addition, we also had a total of 308 filler phrases of verbs with the same inflectional exponents but with different stem vowels (e.g., *ihr spielt*). Furthermore, 231 filler verb phrases in the other combinations of pronouns and suffixes (e.g., *du zahlst*) and 84 filler bare nouns were included. For each participant, a different list was created in which targets and fillers were pseudo-randomized. Because of the relatively large number of items, each list was split into two sublists, which were presented to participants in different sessions.

Procedure

The items were presented visually on a screen of a laptop (Lenovo ThinkPad with intel core i7 and Windows 7), using Articulate Assistant Advanced (Articulate Instruments Ltd., 2012). Participants were instructed to read aloud the items displayed on the screen. Audio signals of their speech were recorded with a microphone (Oktava MK-012) placed about 10 cm away from the mouth (mono audio). The microphone was positioned towards the mouth but slightly lower than the mouth, so that breath does not directly reach the microphone. For an audio interface, we used Focusrite Scarlett Solo 3rd Generation.

At the same time, articulations by participants were recorded with an EchoB ultrasound system (Articulate Instruments Ltd., 2012). The EchoB system included a microconvex 10 mm radius probe by TELEMED, the AAA software, the pulse stretch unit that synchronizes audio signals with ultrasound images, and a probe stabilization headset (UltraFit). Participants were requested to wear the probe stabilization headset, which holds the ultrasound transducer fixed under the chin. The transducer had 64 scanlines, with 842 pixels for each scanline, and with a field of view equal to 92 degrees. Frame rates were 92 frames per second on average.

Before the main session including the target and filler items, participants were instructed to articulate [t] and [k] in addition to swallowing saliva. These data were collected to calculate the position of the hard and soft palate. Cropping points for normalization between speakers were determined in the same way as in Section 3.3.4.

Analysis

Each of the target verb phrases was encountered and produced at least by 10 speakers (min=10, Q1=, Q2=12, Q3=, max=17)[3]. Most (approximately 93%) of the target phrases were spoken by 12 or more speakers. (0.25% by 10 speakers, 6.60% by 11 speakers, 62.69% by 12 speakers, 30.20% by 13 speakers, and 0.25% by 17 speakers).

Recorded ultrasound images were exported as raw data, that is, as a single vector of pixel brightness values. The exported raw data was processed using the python package `pyult` (Saito, 2020). The vector of pixel brightness data was first reshaped as a rectangle, as illustrated above. Subsequently, the reconstructed images were cropped at each side in order to normalize individual differences in oral cavity size (Figure 3.11). These reference points were determined for each participant, based on the middle frame of the recording of [k].

Subsequently, five frames with approximately equal intervals were extracted

[3]Because not all the participants showed up for the second session of the recordings, not all the items could be obtained from every speaker.

from each recording. These five frames corresponded to 0%, 25%, 50%, 75%, and 100% of the stem vowel. For each of the 395 target phrases, the images for a given time point were averaged by speaker.

The procedure above led to the total number of average ultrasound images available for the analysis being $395 \times 5 = 1975$ for each of the target verbal phrasal types, which in all comprised $697,448,365$ pixel values. This large dataset proved to be too large to allow analysis with a single GAM model. We therefore first resized each ultrasound image to 40×40 pixels. This resulted in $395 \times 5 \times 40 \times 40 = 3,160,000$ pixels for each of the verbal phrasal types, comprising $38,960,000$ pixel values. Furthermore, we ran separate GAM models for each individual timestep for each of the four sets of words defined by pronoun and exponent (see Table 3.5), resulting in a total of $5 \times 4 = 20$ models. Although including word as random effect would be preferable, even for these smaller datasets, this proved to be computationally intractable.

For each of the 20 datasets, pixel brightness was modelled as a function of x-coordinate, y-coordinate, vowel duration, word frequency, and interactions of the latter two covariates with the two coordinates. Vowel duration was included to take into account durational differences between the vowels *a* and *a:*, and also because the tongue can move more extensively when more time is available for articulation (Kelso et al., 1985; Kuehn & Moll, 1976). Word frequency was collected from the SdeWac corpus (Faaß & Eckart, 2013) and log-transformed prior to fitting GAMs. The number of basis functions used in the smooth term was set to 20, which we found to be sufficiently large to capture sudden changes in brightness in the ultrasound image, as typically present near the tongue surface. We used the same model formula for all 20 datasets:

```
PixelBrightness ~ s(x, k=20) + s(y, k=20) +
                  ti(x, y, k=20) + s(duration, k=20) +
                  ti(x, duration, k=c(20, 20)) +
                  ti(y, duration, k=c(20, 20)) +
```

```
ti(x, y, duration, k=c(20, 20, 20) +
s(frequency, k=20) +
ti(x, frequency, k=c(20, 20)) +
ti(y, frequency, k=c(20, 20)) +
ti(x, y, frequency, k=c(20, 20, 20))
```

Since we have 20 models with 11 smooth terms each, we used a Bonferroni corrected alpha level $0.01/(20 \times 11) \approx 0.00005$.

3.4.2 Results

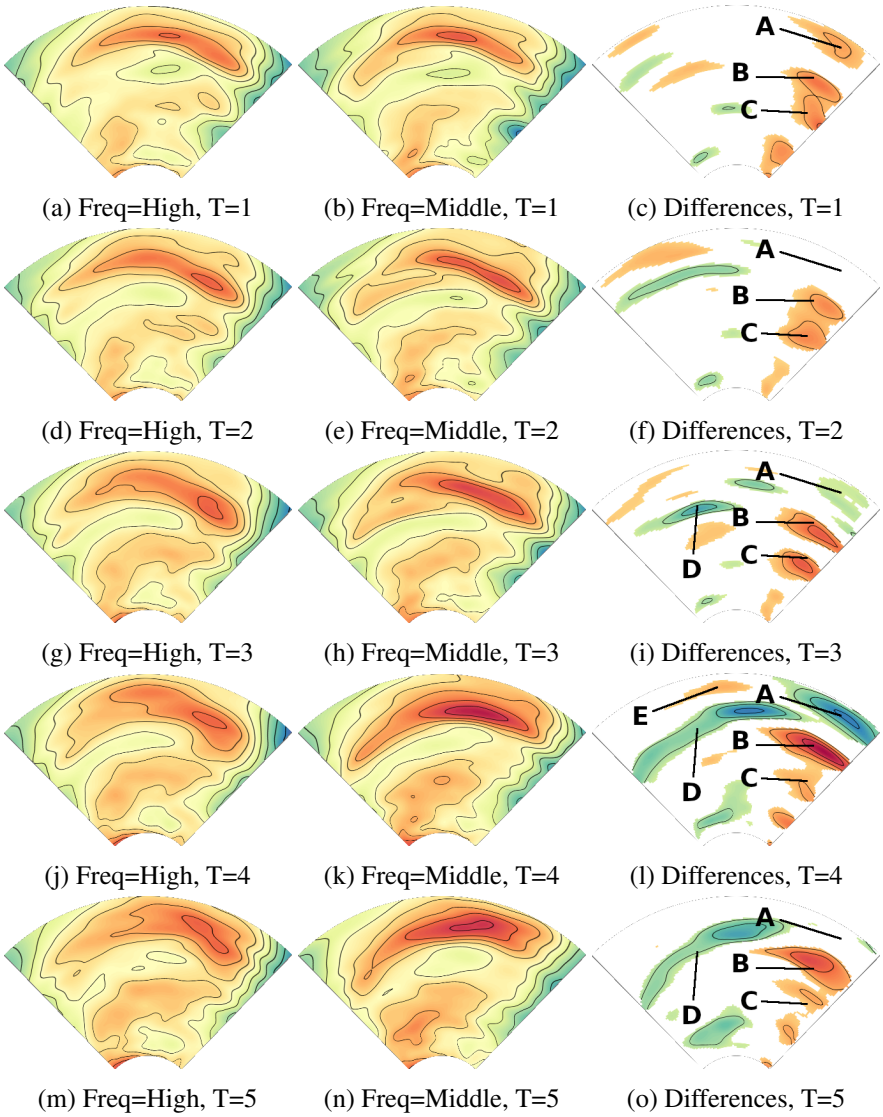
All the smooths were significant at $\alpha = 0.00005$. Summaries of each model are provided in Appendix 3.D.

The suffix-[t] condition

We discuss the results for the condition "iː-t" first, where the pronoun ends with [-iː] and the suffix is [-t]. The predicted ultrasound images for this condition are listed in Figure 3.13. The five plots in the leftmost column represent high frequency words. Those in the central column represent middle frequency words. The rightmost column lists the five difference plots for each time step. Time steps are represented on rows. The top row is the first time step, namely the onset of the target vowel [a(ː)]. The third row is the middle of the vowel. The bottom row is the offset of the vowel. For the predicted ultrasound images in the left and central columns, warmer colors represent brighter pixel values.

In all the GAM models in this case study, frequency was included as a continuous variable. However, just for illustrative purpose, in Figure 3.13 and all the other figures that follow, we refer to 50% and 90% quantiles of the frequency variable as "high" and "mid" frequencies.

The difference plots in the rightmost column were obtained by subtracting the images in the central column from those in the leftmost column. Therefore, warmer colors in the difference plots indicate that pixels are brighter in the high frequency



(a) Freq=High, T=1 (b) Freq=Middle, T=1 (c) Differences, T=1

(d) Freq=High, T=2 (e) Freq=Middle, T=2 (f) Differences, T=2

(g) Freq=High, T=3 (h) Freq=Middle, T=3 (i) Differences, T=3

(j) Freq=High, T=4 (k) Freq=Middle, T=4 (l) Differences, T=4

(m) Freq=High, T=5 (n) Freq=Middle, T=5 (o) Differences, T=5

Figure 3.13: Predicted ultrasound images for words preceded by *sie* and ending with the exponent *-t*. High frequency words are presented in the left column, middle frequency words are presented in the center column, and the corresponding difference surfaces in the right column.

condition, compared to the middle frequency condition. Colder colors indicate the opposite, namely brighter pixels in the middle frequency condition than in the high frequency condition.

Differences between the high-frequency and mid-frequency words that are visible to the eye are the lower position of the front of the tongue for higher-frequency

words at timesteps 3 and 4, and to some extent timestep 5. Furthermore, at later timesteps, there is greater uncertainty about the position of the tongue surface, as indicated by the greater areas with darker red. This uncertainty appears to be somewhat less for the mid-frequency words at timesteps 4 and 5, especially for the tongue root.

The difference surfaces in the right column provide further information. At the first timestep, the front of the tongue is more likely to be positioned in areas A and B for high-frequency words. Area C probably is close to the air pocket below the tongue tip. At the second timestep, area B is again more likely for high-frequency words, but for area A, no difference is present. At timestep 3, area A begins to show a preference for mid-frequency words, and by timestep 4, this preference is very clearly present. At this timestep, the front of the tongue is positioned higher for mid-frequency words, and lower for high-frequency words. At vowel offset, only area B remains as area of preference for high-frequency words.

The difference surface at timestep 4 also clarifies that the tongue back and tongue root are more likely to be in area D for mid-frequency words, whereas for high-frequency words, there center/back of the tongue is more likely to be located higher in area E.

In summary, for higher-frequency words, the front of the tongue is lowered more, and the center/back of the tongue raised more, compared to mid-frequency words, with the greatest differences emerging at T=4.

Figure 3.14 shows, for the *ihr -t* condition, the predicted ultrasound images for high frequency words (left column) and middle frequency words (central column) from the onset (top row) to the offset (bottom row) of the stem vowel. The right column of Figure 3.14 contains the difference surfaces between the predicted ultrasound images of high and middle frequency words. As for the *sie -t* condition, lowering of the tongue tip is visible during timesteps 2, 3, and 4. Unlike the *sie -t* condition, the variance in tongue positions is very similar across all time steps for both high and medium frequency words. With respect to the differences between

the high and middle frequency conditions, the difference surfaces show that these are much reduced. There is some evidence that the tongue position is higher for the high frequency words (area A), starting with the tongue tip at the first time steps, and moving to the tongue body at later time steps. It is only at the first time step that there is some evidence for the front of the tongue being located further down (area B). By the end of the vowel, the tongue back and the tongue root appear to be positioned somewhat lower for the medium frequency words (area C). For these words, there is more reflection from tongue fat in the tongue root (area D), possibly due to a more fronted position of the tongue tip (area E).

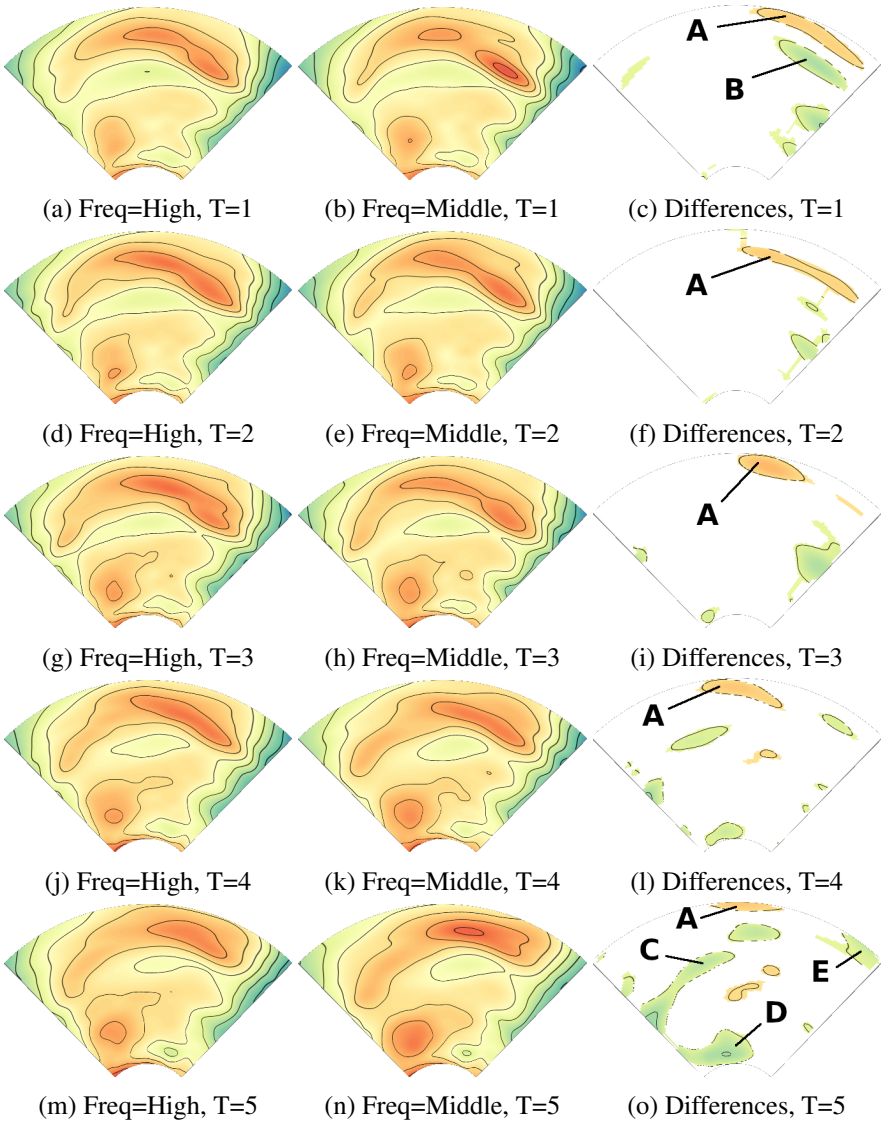
Comparing the fitted surfaces at T=1 for the *sie -t* and the *ihr -t* condition, the position of the tongue front appears to be slightly lower for the *ihr -t* condition, suggesting coarticulation with the preceding vowel, which has a lower point of articulation for the latter condition ([ɐ] vs. [iː]).

The suffix-[(ə)n] condition

Figures 3.15-3.16 present the results for the *sie -n* and *wir -n* conditions respectively.

In the *sie -n* condition, some lowering of the front of the tongue is visible at the second time step. At later time steps, the front of the tongue is found at increasingly high positions especially for middle frequency words. The difference surfaces show that, at vowel onset, higher frequency words have a higher position of the front of the tongue than middle frequency words (areas A and B). This difference substantially reduced in the second time step, and is completely absent in the third time step, at the center of the vowel. Subsequently, mid frequency words have a higher position of the tongue back (area C)

In the *wir -n* condition (Figure 3.16), the wider tongue surface regions for the high frequency words indicate greater variability in articulation. For both high and middle frequency words, lowering of the front of the tongue is visible at time steps 2, 3, and 4. The difference surfaces suggest that the higher variance visible for



(a) Freq=High, T=1 (b) Freq=Middle, T=1 (c) Differences, T=1

(d) Freq=High, T=2 (e) Freq=Middle, T=2 (f) Differences, T=2

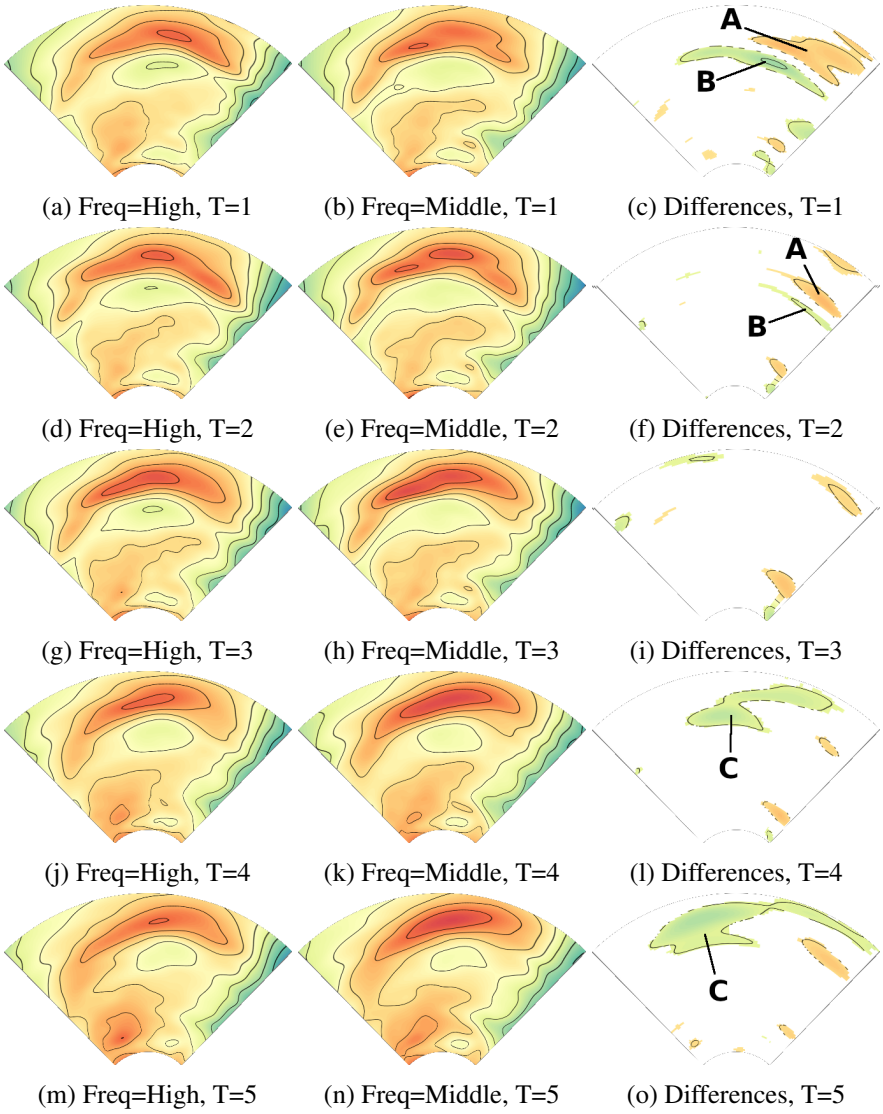
(g) Freq=High, T=3 (h) Freq=Middle, T=3 (i) Differences, T=3

(j) Freq=High, T=4 (k) Freq=Middle, T=4 (l) Differences, T=4

(m) Freq=High, T=5 (n) Freq=Middle, T=5 (o) Differences, T=5

Figure 3.14: Predicted ultrasound images for high frequency (left) and middle frequency words (center) with the pronoun ending [-ɐ] and the suffix being [-t] and differences between the two frequency conditions (right).

high frequency words at many of the time steps is due to two different ways in which speakers realize the vowel, which are jointly represented in the regression surfaces due to aggregation across speakers. Some speakers realize high frequency words with higher tongue positions than the mid frequency words (area A). Other speakers make use of lower positions of the front and the mid of the tongue, and at



(a) Freq=High, T=1 (b) Freq=Middle, T=1 (c) Differences, T=1

(d) Freq=High, T=2 (e) Freq=Middle, T=2 (f) Differences, T=2

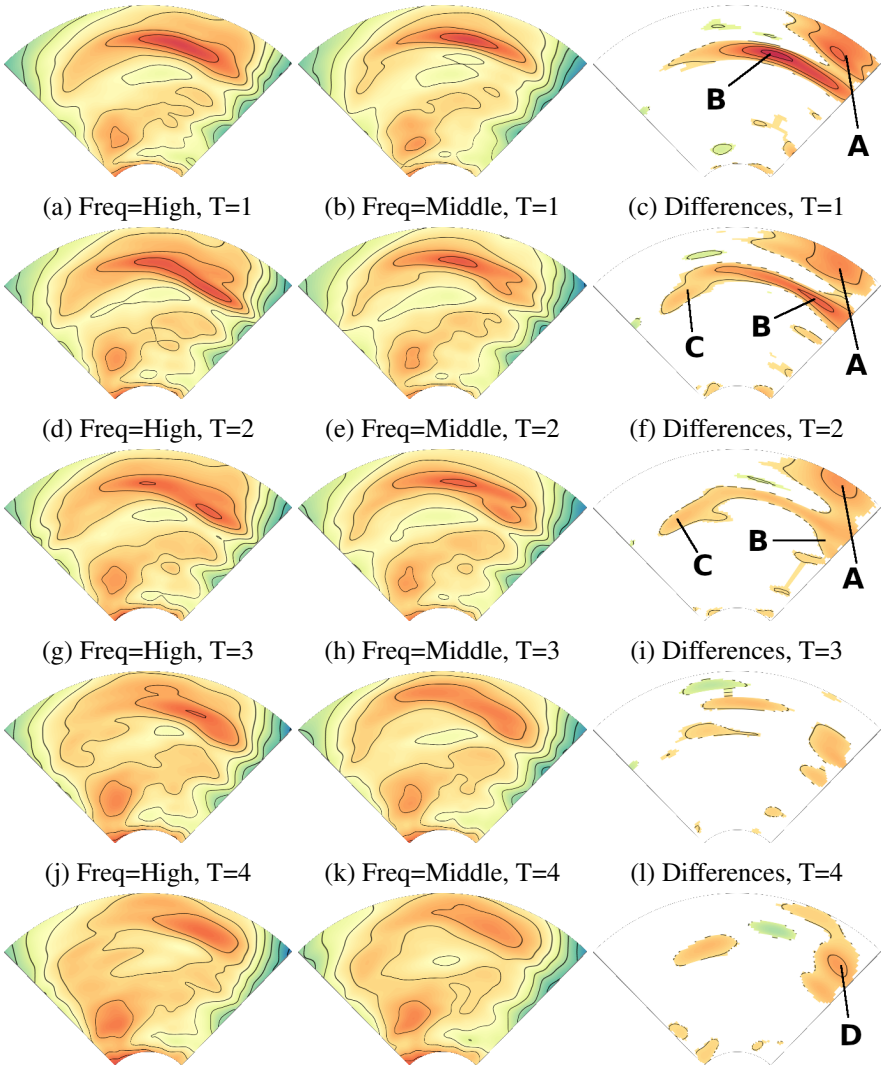
(g) Freq=High, T=3 (h) Freq=Middle, T=3 (i) Differences, T=3

(j) Freq=High, T=4 (k) Freq=Middle, T=4 (l) Differences, T=4

(m) Freq=High, T=5 (n) Freq=Middle, T=5 (o) Differences, T=5

Figure 3.15: Predicted ultrasound images for high frequency (left) and middle frequency words (center) with the pronoun ending [-iː] and the suffix being [-(ə)n] and differences between the two frequency conditions (right).

time steps 2 and 3, even for the tongue back (area C). At time step 4, differences between high and middle frequency words largely disappeared. At vowel offset, higher frequency words show a somewhat more fronted realization.



(a) Freq=High, T=1 (b) Freq=Middle, T=1 (c) Differences, T=1

(d) Freq=High, T=2 (e) Freq=Middle, T=2 (f) Differences, T=2

(g) Freq=High, T=3 (h) Freq=Middle, T=3 (i) Differences, T=3

(j) Freq=High, T=4 (k) Freq=Middle, T=4 (l) Differences, T=4

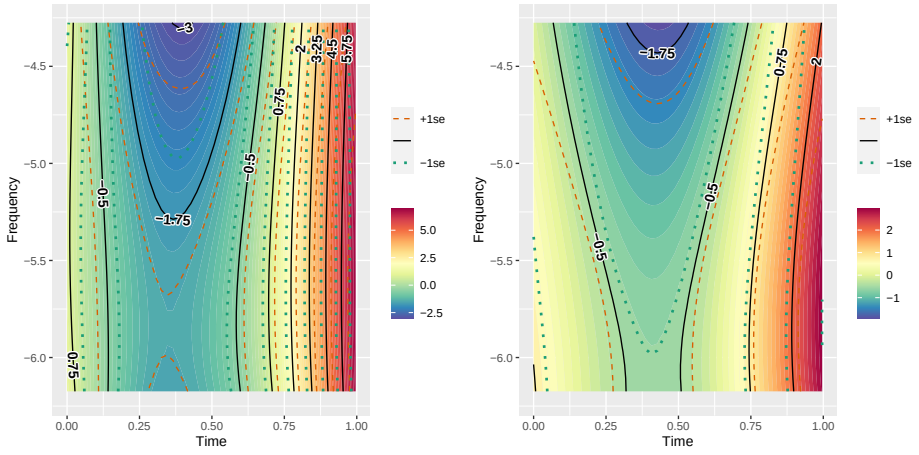
(m) Freq=High, T=5 (n) Freq=Middle, T=5 (o) Differences, T=5

Figure 3.16: Predicted ultrasound images for high frequency (left) and middle frequency words (center) with the pronoun ending [-ɐ] and the suffix being [-t] and differences between the two frequency conditions (right).

3.4.3 Discussion

As mentioned above, this study is in part a replication study of Tomaschek, Tucker, et al. (2018), who used electromagnetic articulography. As can be seen in Figure 3.17, for the *ihr -t* condition, the strongest effect of frequency was present for the tongue tip. In Figure 3.17, the horizontal axis represents time, the vertical

axis represents frequency, and darker colors indicate lower tongue sensor positions. For both the tongue tip and the tongue body sensors, the lowest sensor positions were reached for the highest frequency words. In the present ultrasound study, frequency effects are present for the *ihr -t* condition. However, in this experiment, the tongue front and the tongue back tend to be higher for high frequency words, instead of lower. By contrast, for the *sie -t* condition, which was not considered in Tomaschek, Tucker, et al. (2018), we observed a strong frequency effect that is more similar to that reported in Tomaschek, Tucker, et al. (2018).



(a) Tongue tip (b) Tongue body

Figure 3.17: Tongue tip and body positions for middle to high frequency words in the suffix -t condition found in Tomaschek, Tucker, et al. (2018).

In the *sie -n* condition, in contrast, effects of frequency were not well supported in Tomaschek, Tucker, et al. (2018). In the present ultrasound study, however, we observed a higher position of the tongue front for high frequency words at the onset of the stem vowel, while at the offset of the stem vowel the tongue front was lower for high frequency words. In the *wir -n* condition, which was not considered in Tomaschek, Tucker, et al. (2018), we observed two different kinds of articulatory patterns. For one, the tongue front is higher for high frequency words. For the other, the tongue front is lower for high frequency words.

Across all four conditions (*sie -t*, *ihr -t*, *sie -n*, and *wir -n*), the word-final

exponent contains an alveolar consonant. This consonant may be expected to give rise to raising of the tongue front and the tongue tip by the end of the stem vowel. Expected high positions are indeed visible for all four conditions, but only for middle frequency words. High frequency words show consistently lower tongue positions, typically accompanied by higher positions of the tongue back and the tongue root.[4] This finding fits well with the results reported by Tomaschek, Tucker, et al. (2018), who also observed that higher frequency words were more resistant to coarticulation with upcoming alveolar exponents. As argued by this study, this resistance can be interpreted as evidence for greater articulatory clarity, enabled by more motor experience.

3.5 General discussion

In this paper, we introduced a new statistical method for analyzing ultrasound images. The method is based on the Generalized Additive Model (GAM: Wood, 2017), predicting pixel brightness by x- and y-coordinates, resulting in a predicted wiggly surface for an ultrasound image. These predicted surfaces provide detailed information about the variability in articulation across items. For comparing ultrasound images across experimental conditions, we enriched difference surfaces between these conditions with areas where confidence intervals of the pertinent conditions do not overlap. In this way, combined with a proper Bonferroni correction, the analyst is provided with a guide to where ultrasound images are most likely to be different. The biggest advantage of this method is its ability to include all the information available in an ultrasound image, rather than having to rely on only tongue surface contour lines (as used by, for instance, Aubin & Ménard, 2006; Davidson, 2005, 2006; Dawson et al., 2016; Heyne et al., 2019; Lee-Kim et al., 2013; Ménard et al., 2013; Song et al., 2013; Stolar & Gick, 2013; Strychar-

[4] In the *ihr -t* condition, this is evidenced by the lower position of the tongue back and the tongue root as well as the higher position of the tongue tip for medium frequency words, see areas C and E in Figure 3.14.

czuk & Scobbie, 2016; Turton, 2015). The shadows and the inside of the tongue can also be used as information sources. For example, the movement of the hyoid shadow can be an indication of the advancement of the tongue root. Therefore, this analysis method provides a more holistic view of the tongue and helps to understand what is happening in the oral cavity more clearly. Currently, this method comes with a high computational load. In order to work around this problem, we averaged over subjects.

This method can be applied not only to midsaggital ultrasound images, but also to any other slice of the tongue (e.g., coronal). In principle, extension to 3-D instead of 2-D ultrasound imaging is also possible, in which case higher-dimensional tensor product smooths are required in order to model the interaction of the x, y, and z coordinates with time. Currently, the huge computational load of fitting high-dimensional surfaces to the immense amounts of pixel data generated by 3-D ultrasound imaging, may render a full analysis infeasible for the current algorithms implemented in the **mgcv** package.

To illustrate the GAM method for 2-D ultrasound images, we carried out a replication study with ultrasound, following up on the study of Tomaschek, Tucker, et al. (2018), which made use of electromagnetic articulography. GAM analysis of the ultrasound images revealed similar lower positions of the front of the tongue for high frequency words with [a(ː)] as stem vowel, in general replicating Tomaschek, Tucker, et al. (2018). We observed an effect of frequency for three of the pronoun-suffix conditions in the experiment, the exception being the *wir -n* condition. In this condition, speakers pronounced target phrases in two different ways. This variability may also underlie the fact that Tomaschek, Tucker, et al. (2018) did not find a strong effect of frequency for the suffix *-n* condition.

Lower tongue tip positions for the [a(ː)] vowel indicate clearer articulations. Tomaschek, Tucker, et al. (2018) interpreted this result as indicating enhancement effects of frequency. Clearer articulations for high frequency words are in line with the finding of less co-articulation for adults compared to children (Howson

& Redford, 2019; Nittrouer et al., 1989; Noiray et al., 2019; Sereno et al., 1987; Zharkova et al., 2011). In addition, this finding is also in line with several studies that found enhancement effects of frequency and predictability in morphologically complex words (M. J. Bell et al., 2021; Cohen, 2014; Kuperman et al., 2007; Tomaschek et al., 2021).

These enhancement effects of frequency appear to contradict the reduction effect of frequency, which has been observed repeatedly and well-established (Arnon & Cohen Priva, 2013; Aylett & Turk, 2004, 2006; A. Bell et al., 2009; A. Bell et al., 2002; Dinkin, 2008; Gahl, 2008; Jurafsky et al., 2001; Pluymaekers et al., 2005b; Van Son & Van Santen, 2005; Wright, 2004). One difference between these studies and the present study is that the latter investigates only morphologically complex words, focusing on a stem vowel that is always followed by a morpheme boundary. Importantly, segments before the morpheme boundary have been observed to be enhanced phonetically (Hay, 2007; V. G. Li et al., 2020; Plag & Ben Hedia, 2018; Seyfarth et al., 2017; Smith et al., 2012; Song et al., 2013; Strycharczuk & Scobbie, 2016; Sugahara & Turk, 2009). The study by Tomaschek, Tucker, et al. (2018) and the present replication using ultrasound add to this literature the observation that this enhancement effect appears to be stronger for higher frequency words.

A second difference is that, in the present experiment, target phrases were presented in isolation without context, whereas the reduction effect of frequency is generally understood in terms of syntagmatic predictability (Aylett & Turk, 2006; Jurafsky et al., 2001): more frequent words have a reduced information load in utterances and hence can be reduced to minimize effort (Zipf, 1949). Without any context, the verb phrases in our experiment were syntagmatically unpredictable, which explains why no articulatory reduction took place: all verb phrases were equally (un)informative. This account is in line with Gahl and Baayen (2022), a study tha suggests frequency effects are composed of two opposite forces, namely syntagmatic reduction effects and paradigmatic enhancement effects.

The observed enhancement effects of frequency fit well with the kinematic

practice hypothesis (Tomaschek, Arnold, et al., 2018; Tomaschek, Tucker, et al., 2018), according to which well-practiced routines of tongue movements are executed faster and more clearly (Tomaschek, Arnold, et al., 2018). It is also possible that higher-frequency words are not only executed with better-trained motor skills, but also have stronger form-meaning associations (Gahl & Baayen, 2022). Stronger associations allow words' forms to receive more support from their semantics, which in turn may support phonetically enhanced realization (Chuang et al., 2022; Gahl & Baayen, 2022).

Appendices

3.A Curvature index

Curvature index, denoted i, is defined as

$$
\begin{aligned}
i &= \int_a^b r\,dx \\
r &= \frac{(1+y'^2)^{\frac{3}{2}}}{(y'')} \\
y &= ax_7 + bx_6 + cx_5 + dx_4 + ex_5 + fx_2 + gx + h
\end{aligned}
$$

where x and y are coordinates of a fitted tongue contour line. y' and y'' represent the first and second derivatives of y. a and b are the both ends of the tongue surface contour (Stolar & Gick, 2013).

3.B Discrete Fourier Transform (DFT)

Suppose that we have n data points, which are assumed to be sampled from a continuous periodic function, i.e., $f_0, f_1, f_2, ..., f_{n-1}$. These data points are defined in the time domain. Discrete Fourier Transform maps these data points onto corre-

sponding data points in the frequency domain, i.e., $\hat{f}_0, \hat{f}_1, \hat{f}_2, ..., \hat{f}_{n-1}$. The coefficient of kth frequency (i.e., $\hat{f}_k$) is then defined by the following equation with i an imaginary number (i.e., $i = \sqrt{-1}$):

$$\hat{f}_k = \sum_{j=0}^{n-1} f_j e^{-i2\pi jk/n} \tag{3.1}$$

This equation can be rewritten as below, using Euler's formula $e^{ix} = \cos x + i\sin x$. As indicated by the equation, DFT approximates the original function by a sum of cosine and sine functions.

$$\sum_{j=0}^{n-1} f_j e^{-i2\pi jk/n} = \sum_{j=0}^{n-1} f_j(\cos(2\pi jk/n) - i\sin(2\pi jk/n)) \tag{3.2}$$

The DFT can also be expressed as below, where $\omega_n = e^{-i2\pi/n}$. This equation shows more explicitly that DFT is a linear mapping from the time domain to the frequency domain.

$$\begin{bmatrix} \hat{f}_0 \\ \hat{f}_1 \\ \hat{f}_2 \\ \vdots \\ \hat{f}_{n-1} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega_n & \omega_n^2 & \cdots & \omega_n^{n-1} \\ 1 & \omega_n^2 & \omega_n^4 & \cdots & \omega_n^{2(n-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega_n^{n-1} & \omega_n^{2(n-1)} & \cdots & \omega_n^{(n-1)^2} \end{bmatrix} \begin{bmatrix} f_0 \\ f_1 \\ f_2 \\ \vdots \\ f_{n-1} \end{bmatrix} \tag{3.3}$$

3.C Items

Table 3.C.1: The target phrases adopted in the case study.

wir	baden	wir	waten	sie	planen	sie	rast
sie	badet	sie	watet	sie	wagt	ihr	rast
wir	bahnen	wir	zahlen	ihr	wagt	wir	rasen
sie	bahnt	sie	zahlt	wir	wagen	sie	rasen
sie	bat	wir	zahnen	sie	wagen	sie	masst
wir	baten	sie	zahnt	sie	klagt	ihr	masst
sie	baten	sie	ahmt	ihr	klagt	wir	massen
ihr	batet	ihr	ahmt	wir	klagen	sie	massen
wir	blasen	wir	ahmen	sie	klagen	sie	grast
wir	fahnden	sie	ahmen	wir	raten	ihr	grast
sie	fahndet	sie	ahnt	sie	raten	wir	grasen
wir	faseln	ihr	ahnt	ihr	ratet	sie	grasen
sie	faselt	wir	ahnen	sie	jagt	sie	nagt
wir	lahmen	sie	ahnen	ihr	jagt	ihr	nagt
sie	lahmt	sie	ahndet	wir	jagen	wir	nagen
wir	mahlen	ihr	ahndet	sie	jagen	sie	nagen
sie	mahlt	wir	ahnden	sie	atmet	sie	labt
wir	mahnen	sie	ahnden	ihr	atmet	ihr	labt
sie	mahnt	sie	sagt	wir	atmen	wir	laben
wir	malen	ihr	sagt	sie	atmen	sie	laben
sie	malt	wir	sagen	wir	braten	sie	prahlt
sie	plagt	sie	sagen	sie	braten	ihr	prahlt
wir	plagen	wir	tragen	ihr	bratet	wir	prahlen
sie	plagen	sie	tragen	sie	strahlt	sie	prahlen
wir	schaben	ihr	tragt	ihr	strahlt	sie	trabt
sie	schabt	wir	fahren	wir	strahlen	ihr	trabt
wir	schaden	sie	fahren	sie	strahlen	wir	traben
sie	schadet	ihr	fahrt	sie	ragt	sie	traben
wir	schlafen	sie	fragt	ihr	ragt	sie	spasst
wir	sassen	ihr	fragt	wir	ragen	ihr	spasst
sie	sassen	wir	fragen	sie	ragen	wir	spassen
ihr	sasst	sie	fragen	wir	graben	sie	spassen
wir	stapeln	wir	schlagen	sie	graben	sie	rahmt
sie	stapelt	sie	schlagen	ihr	grabt	ihr	rahmt
wir	tadeln	ihr	schlagt	sie	tagt	wir	rahmen
sie	tadelt	sie	plant	ihr	tagt	sie	rahmen
wir	tafeln	ihr	plant	wir	tagen	sie	kramt
sie	tafelt	wir	planen	sie	tagen	ihr	kramt

wir	kramen	ihr	asst	sie	fahnden
sie	kramen	wir	lasen	sie	faseln
wir	schalen	sie	lasen	ihr	faselt
sie	schalen	ihr	last	sie	fasten
sie	hakt	wir	stahlen	sie	labern
ihr	hakt	sie	stahlen	sie	lahmen
wir	haken	ihr	stahlt	ihr	lahmt
sie	haken	wir	brachen	sie	mahlen
sie	tratscht	sie	brachen	ihr	mahlt
ihr	tratscht	ihr	bracht	sie	mahnen
wir	tratschen	wir	frassen	ihr	mahnt
sie	tratschen	sie	frassen	sie	malen
sie	gast	ihr	frasst	ihr	malt
ihr	gast	wir	stachen	ihr	plagt
wir	gasen	sie	stachen	sie	schaben
sie	gasen	ihr	stacht	ihr	schabt
wir	sahen	wir	kamen	sie	schaden
sie	sahen	sie	kamen	sie	schlafen
ihr	saht	ihr	kamt	ihr	schlaft
sie	sahnt	wir	nahmen	sie	stapeln
ihr	sahnt	sie	nahmen	sie	tadeln
wir	sahnen	ihr	nahmt	sie	tafeln
sie	sahnen	wir	sprachen	sie	tapern
sie	wahrt	sie	sprachen	sie	waten
ihr	wahrt	ihr	spracht	sie	zahlen
wir	wahren	sie	trat	ihr	zahlt
sie	wahren	wir	traten	sie	zahnen
wir	lagen	sie	traten	ihr	zahnt
sie	lagen	ihr	tratet		
ihr	lagt	sie	ratschen		
wir	gaben	wir	ratschen		
sie	gaben	ihr	ratscht		
ihr	gabt	sie	ratscht		
wir	trafen	sie	baden		
sie	trafen	sie	bahnen		
ihr	traft	ihr	bahnt		
wir	assen	sie	blasen		
sie	assen	ihr	blast		

3.D Model summaries

A. Parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	67.505	0.018	3705.900	< 0.00005
B. Smooth terms	**Edf**	**Ref.df**	**F-value**	**p-value**
s(x)	18.489	18.951	13944.896	< 0.00005
s(y)	18.944	18.999	7103.456	< 0.00005
s(Duration)	18.874	18.996	85.514	< 0.00005
s(Frequency)	18.923	18.997	154.129	< 0.00005
ti(x, y)	301.022	339.278	948.933	< 0.00005
ti(x, Duration)	150.644	186.226	18.742	< 0.00005
ti(y, Duration)	297.941	335.677	17.961	< 0.00005
ti(x, y, Duration)	2206.641	2976.021	6.644	< 0.00005
ti(x, Frequency)	143.709	178.701	11.642	< 0.00005
ti(y, Frequency)	275.992	321.537	11.301	< 0.00005
ti(x, y, Frequency)	1738.910	2374.300	4.692	< 0.00005

Table 3.D.1: Summary of the GAM fitted to the ultrasound image for the *sie -n* condition, at $T = 1$.

A. Parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	68.514	0.018	3852.798	< 0.00005
B. Smooth terms	**Edf**	**Ref.df**	**F-value**	**p-value**
s(x)	18.535	18.958	14829.777	< 0.00005
s(y)	18.950	18.999	8368.386	< 0.00005
s(Duration)	18.906	18.997	119.765	< 0.00005
s(Frequency)	18.931	18.998	167.458	< 0.00005
ti(x, y)	304.617	341.531	1002.922	< 0.00005
ti(x, Duration)	150.722	186.636	20.824	< 0.00005
ti(y, Duration)	305.973	341.278	23.333	< 0.00005
ti(x, y, Duration)	2537.358	3392.393	6.944	< 0.00005
ti(x, Frequency)	143.348	178.267	14.057	< 0.00005
ti(y, Frequency)	272.327	318.410	9.201	< 0.00005
ti(x, y, Frequency)	1618.589	2212.068	3.860	< 0.00005

Table 3.D.2: Summary of the GAM fitted to the ultrasound image for the *sie -n* condition, at $T = 2$.

A. Parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	69.360	0.018	3814.270	< 0.00005
B. Smooth terms	**Edf**	**Ref.df**	**F-value**	**p-value**
s(x)	18.529	18.957	14682.985	< 0.00005
s(y)	18.945	18.999	7809.375	< 0.00005
s(Duration)	18.883	18.996	175.495	< 0.00005
s(Frequency)	18.926	18.997	156.777	< 0.00005
ti(x, y)	302.418	340.269	966.551	< 0.00005
ti(x, Duration)	149.111	184.576	26.562	< 0.00005
ti(y, Duration)	306.026	341.273	30.399	< 0.00005
ti(x, y, Duration)	2582.375	3444.319	7.834	< 0.00005
ti(x, Frequency)	143.107	177.526	15.946	< 0.00005
ti(y, Frequency)	264.984	312.899	8.619	< 0.00005
ti(x, y, Frequency)	1641.322	2252.182	3.737	< 0.00005

Table 3.D.3: Summary of the GAM fitted to the ultrasound image for the *sie -n* condition, at $T = 3$.

A. Parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	70.867	0.020	3613.871	< 0.00005
B. Smooth terms	**Edf**	**Ref.df**	**F-value**	**p-value**
s(x)	18.441	18.942	13640.063	< 0.00005
s(y)	18.923	18.998	5430.417	< 0.00005
s(Duration)	18.849	18.994	260.926	< 0.00005
s(Frequency)	18.896	18.996	174.384	< 0.00005
ti(x, y)	293.244	334.407	876.934	< 0.00005
ti(x, Duration)	149.198	184.877	30.037	< 0.00005
ti(y, Duration)	303.157	338.490	29.793	< 0.00005
ti(x, y, Duration)	2345.501	3164.392	7.234	< 0.00005
ti(x, Frequency)	146.335	181.040	20.332	< 0.00005
ti(y, Frequency)	277.495	320.556	15.702	< 0.00005
ti(x, y, Frequency)	1726.710	2392.433	4.886	< 0.00005

Table 3.D.4: Summary of the GAM fitted to the ultrasound image for the *sie -n* condition, at $T = 4$.

A. Parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	71.312	0.020	3519.466	< 0.00005
B. Smooth terms	**Edf**	**Ref.df**	**F-value**	**p-value**
s(x)	18.351	18.926	14200.805	< 0.00005
s(y)	18.889	18.996	4344.521	< 0.00005
s(Duration)	18.752	18.987	272.148	< 0.00005
s(Frequency)	18.912	18.997	182.108	< 0.00005
ti(x, y)	285.379	328.966	836.196	< 0.00005
ti(x, Duration)	165.382	204.248	28.889	< 0.00005
ti(y, Duration)	304.058	338.849	33.270	< 0.00005
ti(x, y, Duration)	2215.106	3005.886	6.727	< 0.00005
ti(x, Frequency)	155.961	192.266	24.080	< 0.00005
ti(y, Frequency)	265.280	308.800	22.205	< 0.00005
ti(x, y, Frequency)	1677.846	2341.878	5.010	< 0.00005

Table 3.D.5: Summary of the GAM fitted to the ultrasound image for the *sie -n* condition, at $T = 5$.

A. Parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	68.049	0.037	1851.866	< 0.00005
B. Smooth terms	**Edf**	**Ref.df**	**F-value**	**p-value**
s(x)	18.308	18.918	6873.242	< 0.00005
s(y)	18.910	18.994	3247.975	< 0.00005
s(Duration)	18.964	18.999	241.260	< 0.00005
s(Frequency)	18.831	18.987	183.503	< 0.00005
ti(x, y)	285.309	329.004	389.818	< 0.00005
ti(x, Duration)	163.782	200.617	16.094	< 0.00005
ti(y, Duration)	295.992	330.778	14.488	< 0.00005
ti(x, y, Duration)	2148.925	2828.797	5.015	< 0.00005
ti(x, Frequency)	170.914	209.360	16.423	< 0.00005
ti(y, Frequency)	302.834	335.630	17.607	< 0.00005
ti(x, y, Frequency)	2053.561	2732.973	4.673	< 0.00005

Table 3.D.6: Summary of the GAM fitted to the ultrasound image for the *sie -t* condition, at $T = 1$.

A. Parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	69.528	0.027	2535.116	< 0.00005
B. Smooth terms	**Edf**	**Ref.df**	**F-value**	**p-value**
s(x)	18.404	18.936	7928.390	< 0.00005
s(y)	18.928	18.996	4230.020	< 0.00005
s(Duration)	18.934	18.998	171.740	< 0.00005
s(Frequency)	18.851	18.989	167.968	< 0.00005
ti(x, y)	295.504	335.830	420.687	< 0.00005
ti(x, Duration)	165.707	202.718	15.379	< 0.00005
ti(y, Duration)	296.266	331.771	11.827	< 0.00005
ti(x, y, Duration)	2285.896	3009.217	4.007	< 0.00005
ti(x, Frequency)	180.219	220.037	15.636	< 0.00005
ti(y, Frequency)	307.451	339.200	13.401	< 0.00005
ti(x, y, Frequency)	2144.531	2844.486	3.720	< 0.00005

Table 3.D.7: Summary of the GAM fitted to the ultrasound image for the *sie -t* condition, at $T = 2$.

A. Parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	70.965	0.036	1972.437	< 0.00005
B. Smooth terms	**Edf**	**Ref.df**	**F-value**	**p-value**
s(x)	18.413	18.937	7861.766	< 0.00005
s(y)	18.915	18.994	4185.174	< 0.00005
s(Duration)	18.916	18.997	151.300	< 0.00005
s(Frequency)	18.802	18.982	124.208	< 0.00005
ti(x, y)	297.906	337.369	397.921	< 0.00005
ti(x, Duration)	173.661	212.028	17.656	< 0.00005
ti(y, Duration)	288.773	326.122	12.678	< 0.00005
ti(x, y, Duration)	2213.972	2921.024	4.473	< 0.00005
ti(x, Frequency)	175.702	215.175	16.299	< 0.00005
ti(y, Frequency)	301.531	335.871	12.879	< 0.00005
ti(x, y, Frequency)	2171.929	2877.893	3.919	< 0.00005

Table 3.D.8: Summary of the GAM fitted to the ultrasound image for the *sie -t* condition, at $T = 3$.

A. Parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	72.292	0.032	2292.622	< 0.00005
B. Smooth terms	**Edf**	**Ref.df**	**F-value**	**p-value**
s(x)	18.387	18.934	8091.680	< 0.00005
s(y)	18.871	18.992	3186.284	< 0.00005
s(Duration)	18.930	18.997	152.870	< 0.00005
s(Frequency)	18.819	18.985	101.628	< 0.00005
ti(x, y)	279.702	324.812	424.201	< 0.00005
ti(x, Duration)	166.068	203.260	18.266	< 0.00005
ti(y, Duration)	302.261	334.332	17.414	< 0.00005
ti(x, y, Duration)	2077.898	2758.937	6.155	< 0.00005
ti(x, Frequency)	177.489	217.698	17.829	< 0.00005
ti(y, Frequency)	300.743	334.641	18.595	< 0.00005
ti(x, y, Frequency)	2257.327	2985.417	5.759	< 0.00005

Table 3.D.9: Summary of the GAM fitted to the ultrasound image for the *sie -t* condition, at $T = 4$.

A. Parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	72.502	0.030	2451.159	< 0.00005
B. Smooth terms	Edf	Ref.df	F-value	p-value
s(x)	18.302	18.915	8236.954	< 0.00005
s(y)	18.808	18.987	2377.272	< 0.00005
s(Duration)	18.937	18.998	183.732	< 0.00005
s(Frequency)	18.863	18.991	132.695	< 0.00005
ti(x, y)	282.474	327.145	418.480	< 0.00005
ti(x, Duration)	188.517	228.513	19.387	< 0.00005
ti(y, Duration)	292.920	326.918	20.354	< 0.00005
ti(x, y, Duration)	1957.506	2624.522	6.525	< 0.00005
ti(x, Frequency)	187.828	229.636	15.948	< 0.00005
ti(y, Frequency)	296.812	332.109	25.700	< 0.00005
ti(x, y, Frequency)	2223.379	2976.094	5.671	< 0.00005

Table 3.D.10: Summary of the GAM fitted to the ultrasound image for the *sie -t* condition, at $T = 5$.

A. Parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	68.294	0.039	1748.140	< 0.00005
B. Smooth terms	Edf	Ref.df	F-value	p-value
s(x)	18.412	18.940	7930.030	< 0.00005
s(y)	18.932	18.996	4276.388	< 0.00005
s(Duration)	18.170	18.575	139.104	< 0.00005
s(Frequency)	18.812	18.985	156.147	< 0.00005
ti(x, y)	291.073	333.117	498.757	< 0.00005
ti(x, Duration)	149.976	186.556	14.023	< 0.00005
ti(y, Duration)	303.865	338.046	15.310	< 0.00005
ti(x, y, Duration)	1806.133	2430.689	4.213	< 0.00005
ti(x, Frequency)	146.087	180.858	13.541	< 0.00005
ti(y, Frequency)	301.572	336.555	12.926	< 0.00005
ti(x, y, Frequency)	1905.217	2544.983	4.205	< 0.00005

Table 3.D.11: Summary of the GAM fitted to the ultrasound image for the *wir -n* condition, at $T = 1$.

A. Parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	69.644	0.047	1476.746	< 0.00005
B. Smooth terms	**Edf**	**Ref.df**	**F-value**	**p-value**
s(x)	18.526	18.960	8803.581	< 0.00005
s(y)	18.950	18.998	5222.781	< 0.00005
s(Duration)	18.449	18.779	179.242	< 0.00005
s(Frequency)	18.814	18.985	188.665	< 0.00005
ti(x, y)	301.606	339.904	526.560	< 0.00005
ti(x, Duration)	160.907	200.469	12.058	< 0.00005
ti(y, Duration)	293.777	332.657	10.719	< 0.00005
ti(x, y, Duration)	1813.760	2438.876	3.529	< 0.00005
ti(x, Frequency)	155.611	191.954	13.685	< 0.00005
ti(y, Frequency)	276.549	318.391	8.884	< 0.00005
ti(x, y, Frequency)	1887.875	2539.102	3.237	< 0.00005

Table 3.D.12: Summary of the GAM fitted to the ultrasound image for the *wir -n* condition, at $T = 2$.

A. Parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	71.244	0.047	1522.831	< 0.00005
B. Smooth terms	**Edf**	**Ref.df**	**F-value**	**p-value**
s(x)	18.536	18.960	9041.976	< 0.00005
s(y)	18.951	18.998	5477.807	< 0.00005
s(Duration)	18.448	18.779	178.046	< 0.00005
s(Frequency)	18.811	18.985	156.850	< 0.00005
ti(x, y)	303.327	340.876	520.514	< 0.00005
ti(x, Duration)	159.003	198.196	11.585	< 0.00005
ti(y, Duration)	295.639	333.811	12.557	< 0.00005
ti(x, y, Duration)	1912.708	2577.254	3.946	< 0.00005
ti(x, Frequency)	161.398	199.185	11.790	< 0.00005
ti(y, Frequency)	272.701	315.852	8.504	< 0.00005
ti(x, y, Frequency)	1820.907	2451.486	3.284	< 0.00005

Table 3.D.13: Summary of the GAM fitted to the ultrasound image for the *wir -n* condition, at $T = 3$.

A. Parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	72.938	0.055	1324.644	< 0.00005
B. Smooth terms	Edf	Ref.df	F-value	p-value
s(x)	18.487	18.953	8442.756	< 0.00005
s(y)	18.904	18.995	3659.194	< 0.00005
s(Duration)	18.829	18.972	164.767	< 0.00005
s(Frequency)	18.788	18.987	169.737	< 0.00005
ti(x, y)	290.324	333.318	488.838	< 0.00005
ti(x, Duration)	146.156	182.942	13.618	< 0.00005
ti(y, Duration)	297.468	333.002	17.336	< 0.00005
ti(x, y, Duration)	1872.196	2512.378	5.400	< 0.00005
ti(x, Frequency)	155.153	192.376	13.489	< 0.00005
ti(y, Frequency)	290.635	329.029	14.696	< 0.00005
ti(x, y, Frequency)	1986.846	2670.292	4.901	< 0.00005

Table 3.D.14: Summary of the GAM fitted to the ultrasound image for the *wir -n* condition, at $T = 4$.

A. Parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	73.451	0.067	1088.345	< 0.00005
B. Smooth terms	Edf	Ref.df	F-value	p-value
s(x)	18.355	18.928	8417.637	< 0.00005
s(y)	18.842	18.990	2629.753	< 0.00005
s(Duration)	18.945	18.996	177.927	< 0.00005
s(Frequency)	18.807	18.991	181.358	< 0.00005
ti(x, y)	280.965	326.829	443.274	< 0.00005
ti(x, Duration)	151.954	189.231	14.983	< 0.00005
ti(y, Duration)	274.186	314.789	18.562	< 0.00005
ti(x, y, Duration)	1722.641	2352.118	5.473	< 0.00005
ti(x, Frequency)	174.600	214.623	16.562	< 0.00005
ti(y, Frequency)	298.664	334.335	20.048	< 0.00005
ti(x, y, Frequency)	1834.015	2500.661	5.751	< 0.00005

Table 3.D.15: Summary of the GAM fitted to the ultrasound image for the *wir -n* condition, at $T = 5$.

A. Parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	69.263	0.020	3415.147	< 0.00005
B. Smooth terms	**Edf**	**Ref.df**	**F-value**	**p-value**
s(x)	18.427	18.945	11284.596	< 0.00005
s(y)	18.940	18.998	5848.467	< 0.00005
s(Duration)	18.918	18.998	282.603	< 0.00005
s(Frequency)	18.910	18.998	169.082	< 0.00005
ti(x, y)	289.330	331.418	738.939	< 0.00005
ti(x, Duration)	160.655	200.082	16.849	< 0.00005
ti(y, Duration)	299.838	337.183	19.033	< 0.00005
ti(x, y, Duration)	2209.780	2994.631	7.109	< 0.00005
ti(x, Frequency)	152.209	188.502	9.949	< 0.00005
ti(y, Frequency)	286.054	326.488	11.542	< 0.00005
ti(x, y, Frequency)	1661.307	2266.915	3.924	< 0.00005

Table 3.D.16: Summary of the GAM fitted to the ultrasound image for the *ihr -t* condition, at $T = 1$.

A. Parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	70.450	0.020	3531.458	< 0.00005
B. Smooth terms	**Edf**	**Ref.df**	**F-value**	**p-value**
s(x)	18.478	18.953	11004.234	< 0.00005
s(y)	18.951	18.998	6691.142	< 0.00005
s(Duration)	18.909	18.998	202.787	< 0.00005
s(Frequency)	18.882	18.997	172.550	< 0.00005
ti(x, y)	299.900	338.690	739.345	< 0.00005
ti(x, Duration)	158.240	196.908	19.981	< 0.00005
ti(y, Duration)	295.960	335.449	19.105	< 0.00005
ti(x, y, Duration)	2472.935	3316.275	7.033	< 0.00005
ti(x, Frequency)	143.138	178.053	9.702	< 0.00005
ti(y, Frequency)	277.622	319.841	11.058	< 0.00005
ti(x, y, Frequency)	1751.961	2388.756	3.752	< 0.00005

Table 3.D.17: Summary of the GAM fitted to the ultrasound image for the *ihr -t* condition, at $T = 2$.

A. Parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	71.979	0.020	3520.357	< 0.00005
B. Smooth terms	Edf	Ref.df	F-value	p-value
s(x)	18.459	18.948	10394.568	< 0.00005
s(y)	18.943	18.998	6254.787	< 0.00005
s(Duration)	18.895	18.997	234.111	< 0.00005
s(Frequency)	18.857	18.996	158.517	< 0.00005
ti(x, y)	296.942	336.742	704.106	< 0.00005
ti(x, Duration)	152.959	190.244	27.287	< 0.00005
ti(y, Duration)	299.519	337.403	26.949	< 0.00005
ti(x, y, Duration)	2477.328	3343.370	7.701	< 0.00005
ti(x, Frequency)	137.617	171.572	10.652	< 0.00005
ti(y, Frequency)	288.315	327.473	13.015	< 0.00005
ti(x, y, Frequency)	1807.037	2461.390	4.221	< 0.00005

Table 3.D.18: Summary of the GAM fitted to the ultrasound image for the *ihr -t* condition, at $T = 3$.

A. Parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	72.936	0.022	3390.891	< 0.00005
B. Smooth terms	Edf	Ref.df	F-value	p-value
s(x)	18.320	18.923	9929.637	< 0.00005
s(y)	18.917	18.997	4271.097	< 0.00005
s(Duration)	18.865	18.996	240.597	< 0.00005
s(Frequency)	18.845	18.995	185.789	< 0.00005
ti(x, y)	285.371	328.815	659.826	< 0.00005
ti(x, Duration)	150.920	187.287	35.869	< 0.00005
ti(y, Duration)	295.435	333.586	31.447	< 0.00005
ti(x, y, Duration)	2385.094	3207.092	8.120	< 0.00005
ti(x, Frequency)	143.572	178.842	11.327	< 0.00005
ti(y, Frequency)	278.325	318.434	17.126	< 0.00005
ti(x, y, Frequency)	1871.948	2559.794	5.374	< 0.00005

Table 3.D.19: Summary of the GAM fitted to the ultrasound image for the *ihr -t* condition, at $T = 4$.

A. Parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	73.665	0.023	3252.502	< 0.00005
B. Smooth terms	**Edf**	**Ref.df**	**F-value**	**p-value**
s(x)	18.246	18.908	10290.045	< 0.00005
s(y)	18.874	18.994	3002.719	< 0.00005
s(Duration)	18.882	18.997	303.469	< 0.00005
s(Frequency)	18.870	18.996	189.700	< 0.00005
ti(x, y)	274.845	320.924	621.766	< 0.00005
ti(x, Duration)	151.897	188.073	36.078	< 0.00005
ti(y, Duration)	293.534	332.516	35.292	< 0.00005
ti(x, y, Duration)	2201.544	2982.905	7.620	< 0.00005
ti(x, Frequency)	152.506	189.062	11.080	< 0.00005
ti(y, Frequency)	267.993	309.394	22.151	< 0.00005
ti(x, y, Frequency)	1836.478	2528.805	5.646	< 0.00005

Table 3.D.20: Summary of the GAM fitted to the ultrasound image for the *ihr -t* condition, at $T = 5$.

Chapter 4

Interaction of frequency and inflectional status

Abstract: High frequency has been associated with phonetic reduction on one hand and phonetic enhancement on the other hand. The present study first looks into the possibility that these opposite frequency effects are at least partially due to the different inflectional status of the items being investigated. Based on tongue position data from a spontaneous speech corpus of German, we found that the stem vowels in inflected words tended to be hyper-articulated (i.e., showing phonetic enhancement), while those in non-inflected words tended to be articulated with more centralized tongue positions. This observed modulation by inflectional status is subsequently investigated from the perspective of distributional semantics. Using Linear Discriminative Learning to study the relation between word embeddings and word forms, we observed that the word-final triphones of inflected words received more support from their embeddings compared to uninflected words. Furthermore, replacement of the two-level factorial predictor for inflectional status with the amount of support received from the semantics, led to substantial improvement in model fit. The implications of these results for models of speech production are discussed.

4.1 Introduction

The consequences of the frequencies with which words are used have been investigated extensively for a wide range of aspects of speech processing, including speech perception and speech production (for an overview, see, e.g., Baayen et al., 2016). And yet it is not entirely clear what frequency and frequency-based measures actually capture.

An influential interpretation of lexical frequency effects in speech production is that higher-frequency words are less informative, and that lower degrees of informativity give rise to higher degrees of phonetic reduction. More probable, and less informative, linguistic units such as high frequency words have been found to undergo more phonetic reduction, resulting in shorter acoustic duration (Arnon & Cohen Priva, 2013; Aylett & Turk, 2004, 2006; A. Bell et al., 2009; A. Bell et al., 2002; Gahl, 2008; Jurafsky et al., 2001; Pluymaekers et al., 2005a, 2005b), more centralized formant realization (Dinkin, 2008; Wright, 2004), and more reduced tongue positions (Lin et al., 2011; Tomaschek, Arnold, et al., 2018; Tomaschek et al., 2013). According to the smooth signal redundancy hypothesis (Aylett & Turk, 2004), the positive correlation between frequency and amount of phonetic reduction arises due to the cognitive system preferring a stable rate of information in the speech signal. To achieve such a smooth signal, less informative words have to be reduced more.

In contrast, Kuperman et al. (2007) found that more probable interfixes between the constituents of Dutch compounds were realized with longer duration, rather than shorter duration. They argued that this unexpected positive correlation of probability and phonetic enhancement is paradigmatic in nature. The more probable an interfix is in the paradigm of compounds sharing the same initial constituent, the more the interfix is enhanced in the speech signal (paradigmatic signal enhancement hypothesis). The enhancement effects of frequency and paradigmatic probability was subsequently replicated for inflectional suffixes (Cohen, 2014) and for stem vowels of inflected verbs (Tomaschek, Tucker, et al., 2018; Tomaschek

et al., 2021).

Why are these opposite directions of frequency effects observed? One possible missing factor is morphological status. The reduction effect of frequency was found when only morphologically simple words are in focus (Lin et al., 2011; Wright, 2004) or when morphologically simple and complex words are not distinguished and mixed (Aylett & Turk, 2004; A. Bell et al., 2009; A. Bell et al., 2002; Dinkin, 2008; Gahl, 2008; Pluymaekers et al., 2005a, 2005b; Tomaschek, Arnold, et al., 2018; Tomaschek et al., 2013). In contrast, the enhancement effect of frequency was found so far exclusively for morphologically complex words (Cohen, 2014; Kuperman et al., 2007; Tomaschek, Tucker, et al., 2018; Tomaschek et al., 2021).

Apart from frequency effects, segments preceding morphological boundaries were found to be acoustically longer (Hay, 2007; Plag & Ben Hedia, 2018; Seyfarth et al., 2017; Smith et al., 2012; Sugahara & Turk, 2009) and articulatorily hyper-articulated (V. G. Li et al., 2020; Smith et al., 2012; Song et al., 2013; Strycharczuk & Scobbie, 2016). These findings suggests that phonetic realizations are enhanced before morphological boundaries. Nevertheless, the effect of the morphological boundary and that of frequency have been investigated so far by and large independently. When frequency effects are investigated, morphological status is controlled or simply ignored. When pre-morpheme-boundary effects are researched, frequency effects are controlled through item selection (Seyfarth et al., 2017; Sugahara & Turk, 2009), statistically (Plag & Ben Hedia, 2018; Smith et al., 2012) or ignored in some cases (Song et al., 2013; Strycharczuk & Scobbie, 2016). Therefore, it is important to clarify to what extent the enhancement effect of frequency and morphological boundary effects are independent, or whether the enhancement effect of frequency is confounded with the effect of the morpheme boundary. This is the first aim of the current study.

The second aim of the current study is to provide an improved understanding of the pre-morpheme-boundary effect. The pre-morpheme-boundary effect has

mainly been explained in terms of the paradigm uniformity hypothesis (Seyfarth et al., 2017). This hypothesis states that members of the same morphological paradigm are similar to each other in phonetic realization. For example, Seyfarth et al. (2017) found longer duration for stems of inflected words (e.g., *frees*), compared to their corresponding morphologically simple words (e.g., *freeze*).

However, an alternative interpretation of the morpheme boundary effect suggests itself within the framework of the discriminative lexicon model (Baayen et al., 2019), a theory that does not require linguistic units such as morphemes, stems, and exponents (Chuang et al., 2020; Gahl & Baayen, 2022; Stein & Plag, 2021). This approach, which integrates distributional semantics into a computational model for lexical processing, predicts that greater support from a word's meaning for its form goes hand in hand with articulatory strengthening. For example, Gahl and Baayen (2022) found that spoken word duration of English homophones was positively correlated with a greater amount of semantic support for the word's form. Well-learned form-meaning relationships are enhanced phonetically, while forms with no support from semantics theoretically predict zero duration (Gahl & Baayen, 2022).

In the light of these findings, we expect that the pre-morpheme-boundary effect may in fact reflect different amounts of semantic support that sublexical word-final form features receive from words' meanings. Providing empirical support for this interpretation is the second aim of the current study.

In the following sections, we first address the interaction of morphological status and frequency. Given the previous studies finding the enhancement effect of frequency for inflected words (Cohen, 2014; Tomaschek et al., 2021), we also focused on inflected words. We expect that the enhancement effect of frequency persists after including in a regression model the interaction between frequency and inflectional status. Given that a majority of studies reporting the reduction effect of frequency mainly inspected morphologically simple words, and that those studies finding enhancement effects of frequency are exclusively investigating mor-

phologically complex words, we also expect that phonetic enhancement is present for inflected words, while reduction is expected for morphologically simple words.

Subsequently, we address the question of the source of the pre-morpheme-boundary effect. To this end, we will first introduce a quantitative measure of semantic support based on the Discriminative Lexicon Model that is a real-valued alternative for the dichotomy between simple and complex words given with the factorial variable of inflectional status. We then evaluate this new measure by investigating its predictivity for tongue trajectories registered with electromagnetic articulography. In the discussion section, we discuss possible implications of our results for the understanding of frequency effects in phonetic realization.

4.2 Frequency and inflectional status

4.2.1 Methods

Data

In order to investigate the interaction of frequency by inflectional status, controlling for segmental similarity is essential. It was, however, impossible to find sufficient pairs of morphologically simple and complex words with identical segments over a reasonably wide range of frequencies (e.g., pairs such as *Macht* 'power' vs. *mach+t* 'makes'). Therefore, we extracted all the words with the same rhyme structure, with the same nucleus, and with the same word-final segment, i.e., [a(ː)(X)t], from the articulography section of the Karl-Eberhard-Corpus of spontaneously spoken southern German (KEC: Arnold & Tomaschek, 2016). Our target vowel is [a(ː)], the long and short low open vowels. The word-final segment, which corresponds to a suffix for inflected verbs, is [t]. To allow an enough number of items to be included, at most one intervening segment was allowed between the target vowel and the word-final [t]. The resulting set of target words comprised inflected or non-inflected words with and without a morphological boundary between the target vowel and the word-final segment. The stems of the target items comprised

not only monomorphemic words but also derived words and compounds. For example, *bemalt* [bəmaːlt] 'paints/painted' consists of a prefix *be-*, a verb *-mal-*, and a inflectional suffix *-t*. *Ausland* [aʊslant] 'foreign country' consists of a prefix *Aus-* and a noun *-land*. The former has a morphological boundary between the target vowel [a(ː)] and the word-final [t], while the latter does not. Under this item selection criteria, we were able to collect 560 word tokens from 88 word types, 48 of which were non-inflected and 40 of which were inflected.

For the selected words, vertical tongue tip and body positions were collected. Since the target vowel is [a(ː)] followed by the word-final [t], the strongest coarticulatory tongue movements were expected for the tongue tip. The tongue body was also included, because a study on coarticulatory tongue movements (Tomaschek, Tucker, et al., 2018) also reported an effect of frequency not only for the tongue tip, but also for the tongue body for words with [a(ː)] and a final [t].

Vertical positions of the tongue tip were distributed mainly within -15 mm and +20 mm from the occlusal plane, which was approximated by having the speaker biting a plastic plate (bite plate) (Arnold & Tomaschek, 2016). In some of the word tokens, measurement errors were so big that registered sensor positions jumped around and did not show any consistent pattern of movements. To deal with these jumping data points, intervals between adjacent data points were calculated within each word token. Extraordinarily large intervals which lay outside 1.5 times the interquartile range were considered to be measurement errors, which amounted to approximately 9.23% of the data points (where each data point pairs time and vertical position), while the number of the word tokens was intact. To avoid that the simple removal of the jumping data points leaves too few data points for the time series of tongue positions for a given word token, the word tokens with less than 4 data points after the removal were removed from the dataset. This resulted in the exclusion of 0.39% of the data points and 6.07% of the word tokens.

These word tokens were distributed as shown in Figure 4.1. The numbers of word tokens spoken by speakers ranged from 1 to 36. The numbers of word types

by speakers ranged from 1 to 15. The mean of the numbers of word tokens spoken by each speaker was 15.029, indicated by the vertical line in Figure 4.1. In Figure 4.1, different word types are illustrated in different colors and marked by "Word ID". For example, speaker "s01" produced the words of interest the most often. Speaker "s35" articulated only one word meeting the criteria of the present study only once, which is the word with ID "w44". Word "w44", *halt*, is listed in the color legend of Figure 4.1.

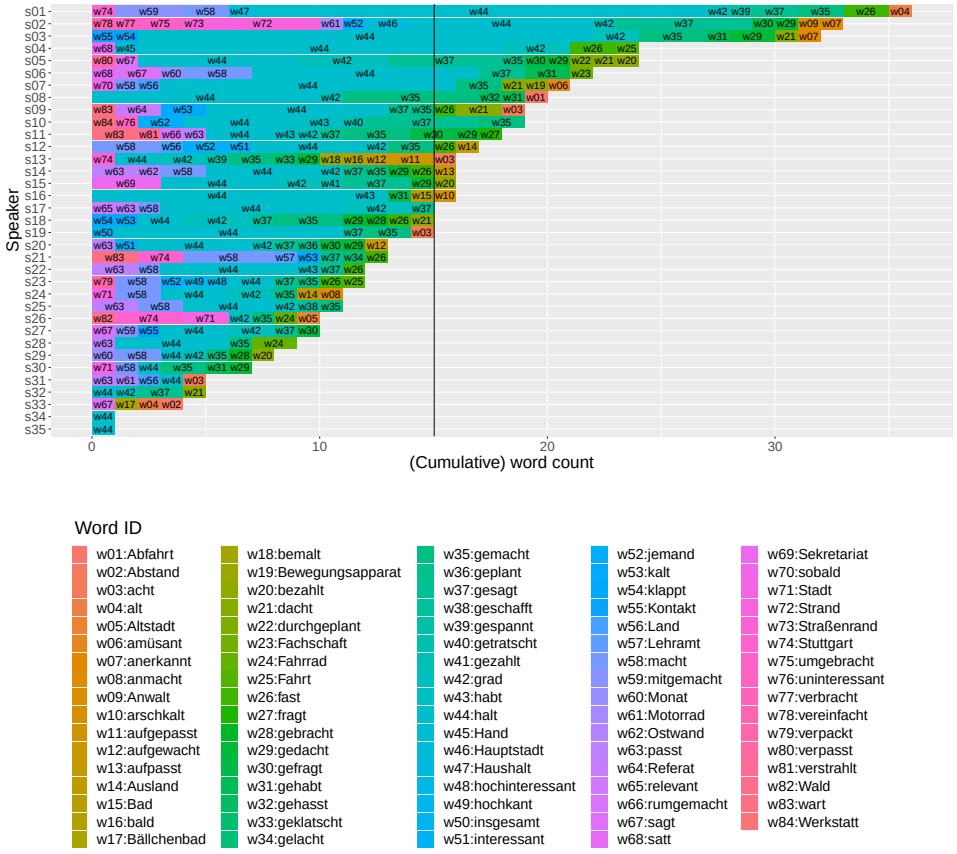


Figure 4.1: The distribution of the words analyzed in the present study across speakers.

Analysis

The tongue movement data during [a(ː)] were fitted with Generalized Additive Mixed-effects Models (GAMMs) (Wood, 2017) for tongue tip movements and tongue body movements separately. In each of the two models, the dependent variable was the vertical position of the tongue tip/body.

Our predictors of interest were time, frequency, and inflectional status (a factor with levels 'non-inflected' vs. 'inflected'). **Word frequency** values were obtained for the target words from the SdeWac corpus (Faaß & Eckart, 2013) and log-transformed prior to the analysis. Log-transformed word frequency was distributed approximately in the same range for non-inflected and inflected words (Figure 4.2). Data points are sparse below log frequencies below 7.

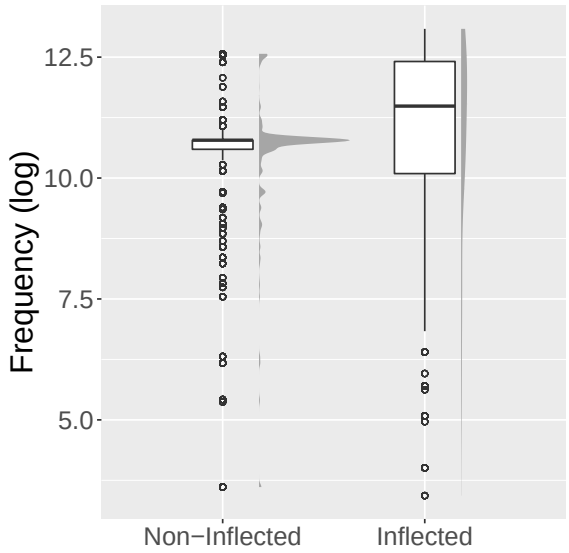


Figure 4.2: Distributions of log-transformed word frequency for non-inflected and inflected words.

Time was normalized between 0 and 1, corresponding to the onset and the offset of the target vowel [a(ː)]. To compensate for the normalization, the target vowel's duration was included as a covariate.

The **duration** of the target vowel was correlated with inflectional condition

($t(368.33) = -4.50$, $p < 0.001$) (Figure 4.3). The vowel [a(ː)] was significantly longer for inflected words, compared to non-inflected words. The longer duration in inflected words is consistent with previous acoustic studies that found similar acoustic lengthening effects in the pre-morpheme-boundary condition (Hay, 2007; V. G. Li et al., 2020; Plag & Ben Hedia, 2018; Seyfarth et al., 2017; Smith et al., 2012; Song et al., 2013; Strycharczuk & Scobbie, 2016; Sugahara & Turk, 2009).

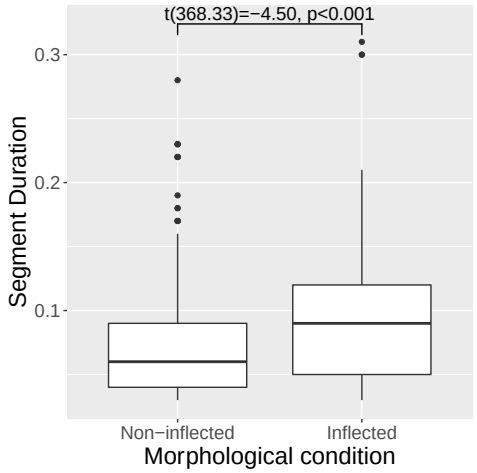


Figure 4.3: The target vowel's duration for inflected and non-inflected words.

In addition, the target vowel's duration was significantly shorter for higher (log) frequency ($r(525) = -0.097$, $p = 0.026$), as illustrated in the left panel in Figure 4.4. The reduction effect of frequency on duration for the present dataset is also in line with previous studies reporting a negative correlation between frequency and segment duration (Aylett & Turk, 2004; A. Bell et al., 2009; A. Bell et al., 2002; Gahl, 2008).

Interestingly, separating the inflected and non-inflected words in the present data, high frequency words turned out to be significantly associated with shorter duration ($r(329) = -0.208$, $p < 0.001$) for non-inflected words, but not for inflected words ($r(194) = -0.040$, $p \approx 0.575$) (see Figure 4.4). A linear model regressing segment duration on frequency, inflectional status, and their interaction supports the presence of the interaction ($t(523) = 2.724, p = 0.007$). This result

suggests that frequency effects play out in different ways for morphologically simple and morphologically complex words.

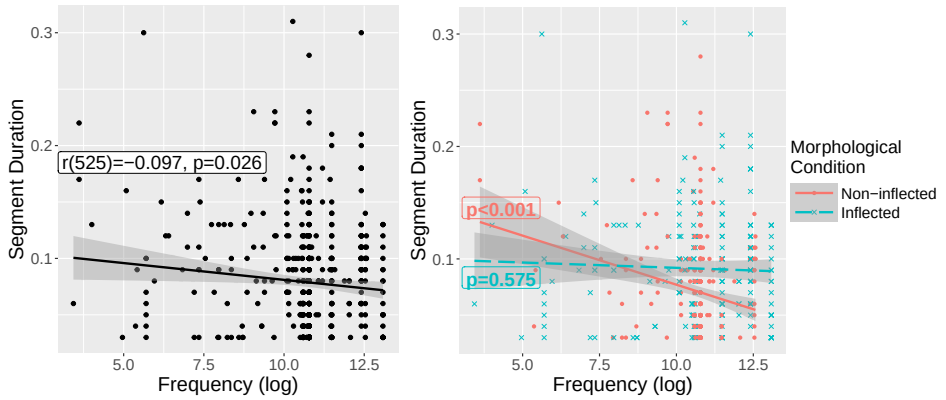


Figure 4.4: Correlation of frequency with the target vowel's duration, aggregating (left plot) and separating (right plot) the inflectional condition.

With respect to random effect factors, speaker and word are two common choices in regression modeling. Although there were differences in the number of tokens uttered by the speakers (see Figure 4.1), including speaker as a random-effect factor was relatively unproblematic. However, as many of the word types were represented by just a single speaker (57%, see Figure 4.5), inclusion of word as random-effect factor is not advisable, as it would lead to an over-specified model (see, e.g., Baayen & Linke, 2020).

As the segments preceding and following the target vowel influence the vowel's articulation, we included random effect factors for these two sets of segments. The distributions of the segments before and after the target vowel are illustrated in Figure 4.6.

Given these predictors, we fitted generalized additive mixed models to the dataset for vertical positions of the tongue tip, and for vertical positions of the tongue body, using the function `bam` of the package `mgcv` (Wood, 2017) in `R` (R Core Team, 2022). As we are interested in the difference surface of time by frequency for simple and complex words, we requested a tensor product smooth with 0/1 coding for inflectional status, as follows:

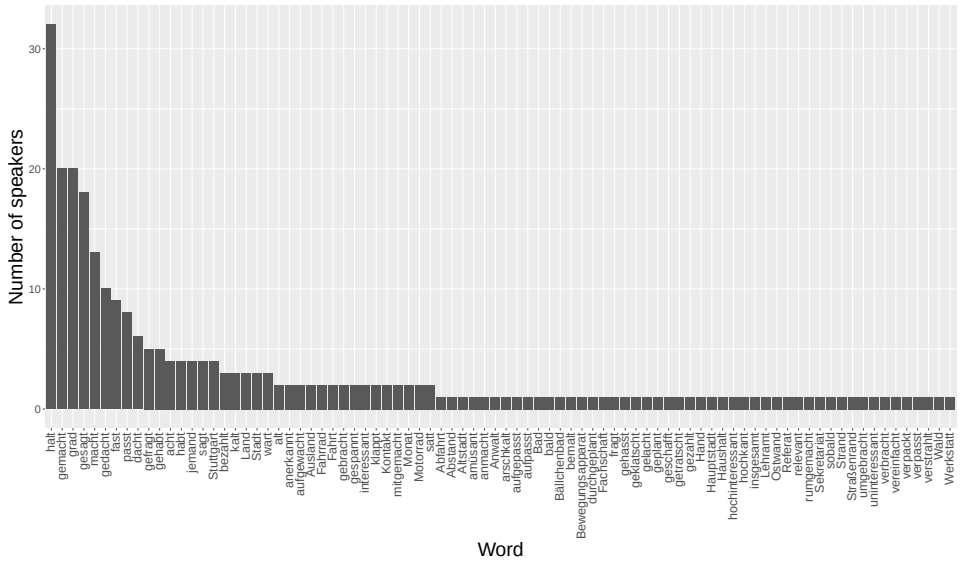


Figure 4.5: Distribution of the word types across the speakers in the present dataset.

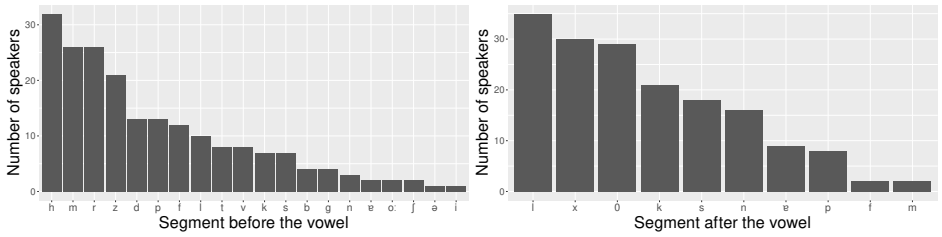


Figure 4.6: Distributions of the segments before and after the vowel of interest.

```
TonguePosition ~ s(Time, Speaker, bs='fs', k=3, m=1) +
                 s(PrevSeg, bs='re', k=3) +
                 s(NextSeg, bs='re', k=3) +
                 s(VowelDuration, k=3) +
                 ti(VowelDuration, Time, k=c(3,3)) +
                 te(Freq, Time, k=c(3,3)) +
                 te(Freq, Time, by=InflStatus, k=c(3,3)) +
                 InflStatus
```

4.2.2 Results

Tongue tip

The fitted GAMM for the tongue tip revealed that articulations of the (stem) vowel [a(ː)] were significantly lower in general for inflected words, compared to non-inflected words, as indicated by the main effect listed in the second row of the upper part of Table 4.1 ($\beta = -4.921, p < 0.001$). Inflectional status interacted with `Time` and `Freq`, as shown in the last row of the lower part of Table 4.1, allowing us to conclude that, for our data, the two regression surfaces are different.

Table 4.1: Summary of the model for the tongue tip.

A. Parametric terms	Estimate	Std.Error	t-value	p-value
Intercept	5.397	2.172	2.485	0.013
Inflected	-4.921	0.467	-10.542	<0.001
B. Smooth terms	**edf**	**Ref.df**	**F**	**p-value**
s(Time, Speaker)	97.067	104.000	666.311	<0.001
s(PrevSeg)	18.170	19.000	251.539	<0.001
s(NextSeg)	8.265	9.000	1112.447	<0.001
s(VowelDuration)	1.002	1.004	32.359	<0.001
ti(Time, VowelDuration)	3.629	3.928	35.911	<0.001
te(Freq, Time)	7.643	7.928	38.272	<0.001
te(Freq, Time):Inflected	7.531	7.900	14.364	<0.001

This interaction is visualized in Figure 4.7. The x-axis represents normalized time, and the y-axis log-transformed frequency. The leftmost and middle panels pertain to non-inflected and inflected words respectively. The rightmost panel displays the difference surface for inflected words. Warmer colors represent higher tongue positions. Since the target vowel is [a(ː)], higher tongue positions indicate articulatory reduction.

For non-inflected words (in the leftmost panel), the tongue tip raises as time proceeds, and reaches its highest elevation at the offset of the vowel. This pattern is present irrespective of frequency. The amount of raising, however, depends on frequency: for the lowest frequency words, the change over time is stronger than

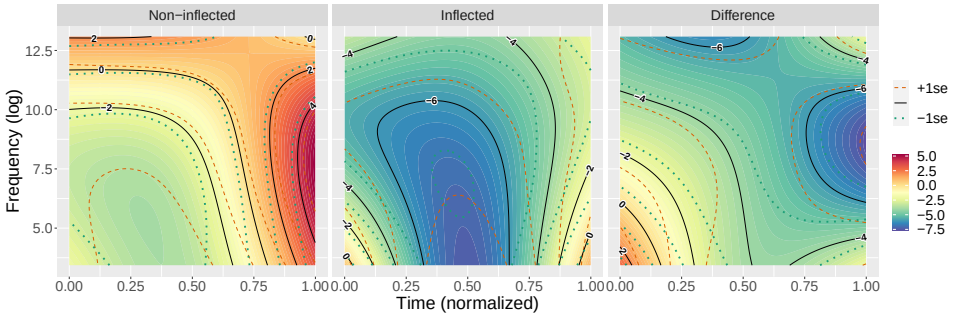


Figure 4.7: Fitted tongue tip height as a function of time and frequency, for non-inflected words (left), inflected words (middle), and the difference surface (right).

for the higher frequency words. Conversely, higher-frequency words are realized with higher tongue positions, most clearly so early in the vowel, and less so near the end of the vowel. In other words, higher-frequency words have higher and flatter trajectories of the tongue tip.

The difference surface is presented in the rightmost panel of Figure 4.7. Addition of this difference surface to the surface of the non-inflected words (i.e., the rightmost panel) results in the predicted surface shown in the middle panel. The middle panel shows that the reduction effect of frequency is retained to some extent also for inflected words. However, tongue tip trajectories for inflected words are overall lower and have a greater lowering of the tongue at the center of the vowel. In addition, the coarticulatory raising of the tongue tip towards the offset of the vowel, which was observed for non-inflected words, is also attenuated substantially for inflected words.

Tongue body

The tongue body also showed a significant main effect of the inflectional condition ($\beta = -1.560, p < 0.001$), as shown in the upper part of Table 4.2. Inflectional status also interacted with frequency and time, albeit to lesser degree compared to the tongue tip model (see the last row of the second half of the table). As can be seen in the left panel of Figure 4.8, non-inflected words are articulated with higher

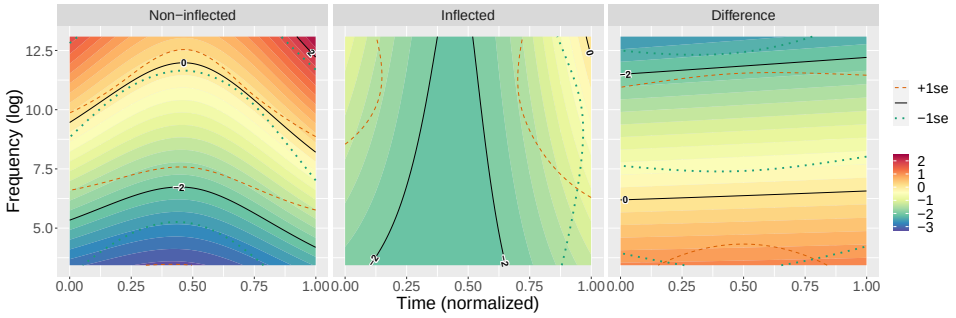


Figure 4.8: Fitted tongue body height as a function of time and frequency, for non-inflected words (left), inflected words (middle), and the difference surface (right).

tongue body positions as frequency increases. These higher tongue positions for higher frequency words are canceled out by the difference surface (the rightmost panel of Figure 4.8), and as a consequence the tongue trajectories of the tongue body are minimal, staying relatively low positions.

Table 4.2: Summary of the model for the tongue body.

A. Parametric terms	Estimate	Std.Error	*t*-value	*p*-value
Intercept	9.388	1.445	6.496	<0.001
Inflected	-1.560	0.410	-3.801	<0.001
B. Smooth terms	**edf**	**Ref.df**	***F***	***p*-value**
s(Time, Speaker)	40.718	104.000	372.099	<0.001
s(PrevSeg)	17.168	19.000	29.062	<0.001
s(NextSeg)	6.797	9.000	112.055	<0.001
s(VowelDuration)	1.019	1.037	0.866	0.363
ti(Time, VowelDuration)	2.904	3.490	5.193	0.001
te(Freq, Time)	4.783	4.967	17.605	<0.001
te(Freq, Time):Inflected	3.018	3.035	3.267	0.020

4.2.3 Interim summary

For both tongue sensors, the GAMMs revealed higher positions for higher frequency words. In addition, the reduction (tongue-raising) effect of frequency was attenuated for inflected words as compared to non-inflected words.

Overall lower tongue positions for inflected words (the main effects, indepen-

dently from frequency and time) are consistent with the paradigm uniformity hypothesis (Seyfarth et al., 2017), according to which phonetic realizations in the pre-morphological-boundary condition should be enhanced. However, this hypothesis does not explain the interaction of the effects of frequency and inflectional status observed for both tongue sensors.

Increases in tongue height hand in hand with increases in frequency reflect articulatory reduction for the [a(ː)]. This effect of frequency dovetails well with the smooth signal redundancy hypothesis (Aylett & Turk, 2004), and is consistent with a number of studies that report reduced phonetic realizations (e.g., Gahl, 2008). However, the smooth signal redundancy hypothesis does not predict attenuation of the reduction effect for inflected words. In the current dataset, we observed a much weaker reduction effect of frequency for inflected words. The attenuated reduction effect of frequency may be due to the opposing pressure enhancing phonetic realizations for clearer articulations. Such opposing enhancement pressure is at least partially in line with the paradigmatic enhancement hypothesis (Kuperman et al., 2007) and the kinematic improvement hypothesis (Tomaschek, Tucker, et al., 2018). However, the absence of such enhancement pressure for non-inflected words remains unaccounted for.

None of these three hypotheses fully explain the articulation patterns observed in the present study for inflected and non-inflected words sufficiently. Therefore, in the next section, we investigate whether the observed patterns of articulation can be explained more precisely in terms of words' inflectional semantics.

4.3 Morpheme boundary or semantics

Several studies framed within the theory of the discriminative lexicon (Baayen et al., 2019; Chuang et al., 2020; Gahl & Baayen, 2022; Stein & Plag, 2021) have reported phonetic enhancement for word forms that are better-supported by their corresponding semantics. For semantically transparent inflected words, strong links

between their forms and meanings are expected, and it is conceivable that these strong links underlie the enhanced articulations reported above.

To test this hypothesis, we will first define a measure of semantic support and show that the semantic measure is correlated with inflectional status. Subsequently, the measure will be used as a predictor for tongue trajectories in a GAM regression model.

4.3.1 Semantic measures derived from the DLM

The discriminative lexicon model (DLM: Baayen et al., 2018; Baayen et al., 2019) is a computational model of lexical processing that works with numerical representations for words' forms and words' semantics. In this study, we represent words' forms with zero/one binary vectors that encode which triphones are present in a word. These vectors are brought together as the row vectors of a word-by-triphone matrix (henceforth $\boldsymbol{C}$). Each word vector (row) in $\boldsymbol{C}$ contains 1 where the triphone in question is contained in the word and 0 otherwise.

Words' meanings are represented by word embeddings. We adopted a pre-trained word2vec model (Müller, 2015) which represents words' semantics with 300 dimensional vectors. These vectors are combined as the row vectors of a word-semantics matrix (henceforth $\boldsymbol{S}$ matrix).

We set up the $\boldsymbol{C}$ (64068, 14404) and $\boldsymbol{S}$ (64068, 300) matrices for all those words in the CELEX database (Baayen et al., 1995) with frequency greater than 0, and for which pre-trained embeddings are available. The DLM posits simple linear mappings between form and meaning matrices. Given $\boldsymbol{C}$ and $\boldsymbol{S}$, a weight matrix $\boldsymbol{F}$, used for modeling comprehension, can be estimated by solving $\boldsymbol{CF} = \boldsymbol{S}$. The obtained $\boldsymbol{F}$ can then be used to estimate a predict semantic matrix $\hat{\boldsymbol{S}}$ by post-multiplication of $\boldsymbol{C}$ by $\boldsymbol{F}$, i.e., $\boldsymbol{CF} = \hat{\boldsymbol{S}}$. Rows of $\hat{\boldsymbol{S}}$ represent predicted word meanings. Conceptually, these are the meanings as understood by the system given the corresponding word-forms. Similarly, a weight matrix $\boldsymbol{G}$ can be estimated for modelling part of the speech production process by solving $\boldsymbol{SG} = \boldsymbol{C}$. The

estimated $\boldsymbol{G}$ maps $\boldsymbol{S}$ onto $\hat{\boldsymbol{C}}$. Rows of $\hat{\boldsymbol{C}}$ are predicted semantic support to word-forms. This method of estimating $\boldsymbol{F}$ and $\boldsymbol{G}$ is called "endstate of learning". For other learning methods implemented for the DLM, see Heitmeier et al. (2022).

Using the endstate-of-learning method in the framework of the discriminative lexicon model, Gahl and Baayen (2022) found that the sum of semantic support from the word to the triphones constituting the word was predictive for word-duration of English homophones (Gahl & Baayen, 2022). Greater semantic support was associated with longer duration (Gahl & Baayen, 2022). For a word *i*, the semantic support for triphone *j* is:

$$\text{SemSup}_{i,j} = \hat{\boldsymbol{C}}_{i,j} \tag{4.1}$$

Let $\mathscr{C}_i$ a set of triphones constituting a word *i*. The sum of semantic supports to all the component triphones of a word *i*, which we call `SemSupWord`, is:

$$\text{SemSupWord}_i = \sum_{k \in \mathscr{C}_i} \hat{C}_{i,k} \tag{4.2}$$

In addition to `SemSupWord`, we considered the triphone centered around the vowel (henceforth the vowel triphone) and the triphone centered around the exponent (henceforth the suffix triphone). Let *v* and *s* denote the indices of the vowel and suffix triphones. The semantic support from a word *i* to its vowel triphone (`SemSupVowel`) and the suffix triphone (`SemSupSuffix`) are defined as

$$\text{SemSupVowel}_i = \hat{\boldsymbol{C}}_{i,v} \tag{4.3}$$

and

$$\text{SemSupSuffix}_i = \hat{\boldsymbol{C}}_{i,s}. \tag{4.4}$$

Along with these measures of semantic support, prediction accuracy of the trained LDL model (i.e., `PredAcc`) was also considered. Prediction accuracy was quantified as the correlation of the predicted and observed (gold-standard) row vectors of

$\hat{C}$ and C. Denoting the i-th row vector of $\hat{C}$ (and C) as $\hat{C}_{i,*}$ ($C_{i,*}$), we have:

$$\text{PredAcc}_i = \text{cor}(\hat{C}_{i,*}, C_{i,*}) \tag{4.5}$$

`PredAcc` was expected to be correlated with the semantic support measures to some extent, especially `SemSupWord`, because well-predicted word-form-vectors should have higher values (only) for their correct component triphones.

In addition to semantic support to form, we also defined another measure that focused on uncertainty among predicted form vectors. Uncertainty among predicted forms (i.e., `UncertProd`) is the product of the correlation of the predicted and observed form vectors and the correlation's rank:

$$\text{UncertProd}_i = \sum_k \big(\text{cor}(\hat{C}_{i,*}, C_{k,*}) \times \text{rank}(\text{cor}(\hat{C}_{i,*}, C_{k,*}))\big). \tag{4.6}$$

The counterpart of this measure for comprehension side is

$$\text{UncertComp}_i = \sum_k \big(\text{cor}(\hat{S}_{i,*}, S_{k,*}) \times \text{rank}(\text{cor}(\hat{S}_{i,*}, S_{k,*}))\big). \tag{4.7}$$

These uncertainty measures are illustrated in Figure 4.1. The left panel presents an example of high uncertainty. The shaded part under the curve represents the uncertainty measure. The predicted word with the highest correlation is found at the right hand side of the plot, with the biggest rank, but many other words are also supported by high correlations. Therefore, even if the most strongly supported word is the correct target word, there are also many other words that are "competitative". In contrast, the right panel shows a case of low uncertainty in prediction. Only one of the possible words is strongly supported with a very high correlation coefficient, and the other words are not well supported. In this example, the word with the greatest rank is supported not only well supported, but also there is little uncertainty about what word is the best candidate.

In addition, the counterpart of semantic support for the comprehension side of the mappings was also considered, i.e. `FuncLoad`. This measure quantifies how

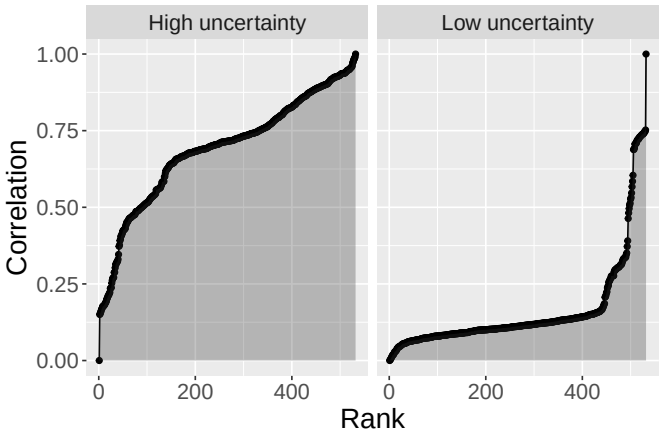


Figure 4.1: Illustration of high and low uncertainty cases.

much triphones help to identify the target word in the comprehension mapping. The `FuncLoad` of a triphone is defined as the correlation of that triphone's row vector in $\boldsymbol{F}$ and the semantic vector of its carrier word in $\hat{\boldsymbol{S}}$. The `FuncLoad` of the j-th triphone to the i-th word is given by

$$\text{FuncLoad}_{j,i} = \text{cor}(\boldsymbol{F}_{j,*}, \hat{\boldsymbol{S}}_{i,*}). \tag{4.8}$$

As for `SemSup`, `FuncLoad` can also be defined for the vowel triphone and the suffix triphone:

$$\text{FuncLoadVowel}_i = \text{cor}(\boldsymbol{F}_{v,*}, \hat{\boldsymbol{S}}_{i,*}), \tag{4.9}$$

$$\text{FuncLoadSuffix}_i = \text{cor}(\boldsymbol{F}_{s,*}, \hat{\boldsymbol{S}}_{i,*}). \tag{4.10}$$

The last measure we considered is the length of a semantic vector (`SemLen`). `SemLen` is simply the L1norm of a semantic vector:

$$\text{SemLen}_i = \sum_j |S_{ij}|. \tag{4.11}$$

4.3.2 Correlation between inflectional status and semantic support measures

How are these semantic measures related to inflectional status? In this section, we first address this question using variable importance measures based on a Random Forest analysis (Breiman, 2001). Subsequently, we look in more detail into how the most important semantic measures pattern with respect to inflectional status.

To this end, all the words from the CELEX database (Baayen et al., 1995) with the stem vowel [a(ː)] and the word-final segment [t], whose frequency was more than 0, were selected. At most one intervening segment between [a(ː)] and [t] was allowed. The resulting dataset comprised 1392 words. Inflectional status was assigned with help of the inflectional information recorded in CELEX. For example, in CELEX, *macht* [maxt] 'makes' is coded as "3SIE,2PIE,rP". The code stands for "third-person singular indicative present (3SIE)", "second-person plural indicative present (2SPIE)", and "imperative plural (rP)". Appendix 4.A provides a complete list of feature bundles and their classification as either "inflected" or "non-inflected".

Variable importance

Inflectional status was entered as the dependent variable in a Random Forest analysis. The number of predictors being considered for a given subsample (i.e., for each split of the tree) was set to three (of the nine semantic measures introduced above), based on a grid-search using the function `train` of the **caret** package (Kuhn, 2021) in `R` (R Core Team, 2022).

The variable importances of the semantic measures are presented in Figure 4.2. `SemSupSuffix` is the best supported predictor for inflectional status, followed by `SemSupVowel`. `SemSupWord` was not as predictive as `SemSupSuffix` and `SemSupVowel`. The good performance of `SemSupSuffix` fits well with the fact that the exponent *-t* has well-defined inflectional meanings, and thus differs from non-inflectional word-final segment such as *-l* (as in *Vogel*, *Ball*).

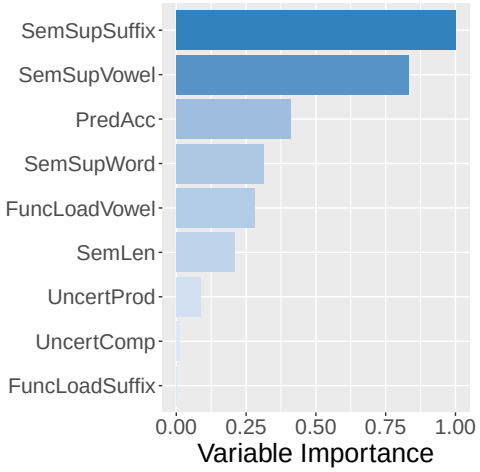


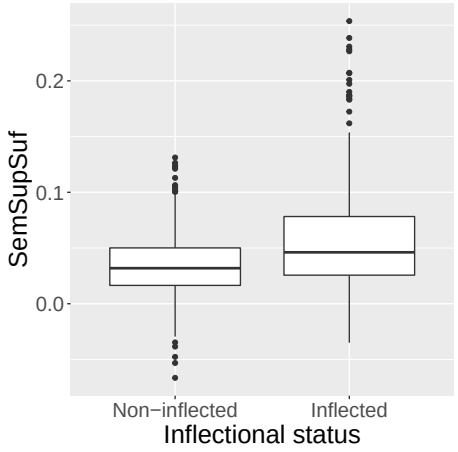
Figure 4.2: Variable importances of the semantic measures.

In what follows, we focus on three of the best-supported measures, which are namely `SemSupSuffix`, `SemSupVowel`, and `SemSupWord`, and look into how they are correlated with inflectional status. Since `PredAcc` is highly correlated with `SemSupWord` ($r = 0.857$), this predictor was not considered further.

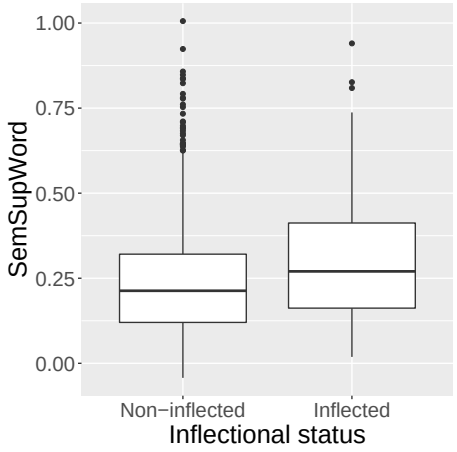
Predicting inflectional status with semantic measures

Inflected words had significantly higher values of semantic support for the suffix and the entire word (U=152201, N1=922, N2=470, $p < 0.0001$ for `SemSupSuffix`; U=172134, N1=922, N2=470, $p < 0.0001$ for `SemSupWord`), as illustrated in Figures 4.3a and 4.3b respectively. By contrast, inflected words were associated with significantly lower `SemSupVowel` (U=266563, N1=922, N2=470, $p < 0.0001$) as can be seen in Figure 4.3c.

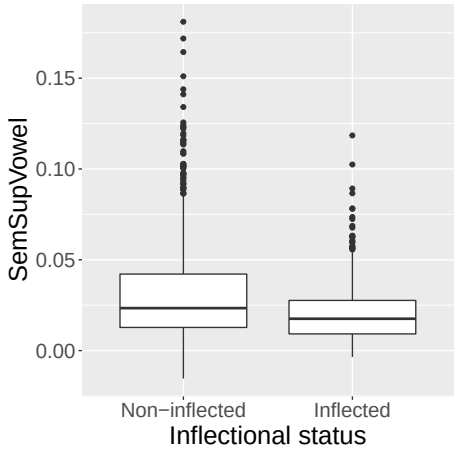
Subsequently, we fitted a logistic regression model, in which the dependent variable was inflectional status. The goal of the logistic model was to predict the probability of a word being inflected. The predictors were `SemSupSuffix`, `SemSupVowel`, and `SemSupWord`. Due to moderate correlations between the three semantic support measures, three logistic regression models were fitted for each of the three semantic measures. Each of the three models showed that the semantic



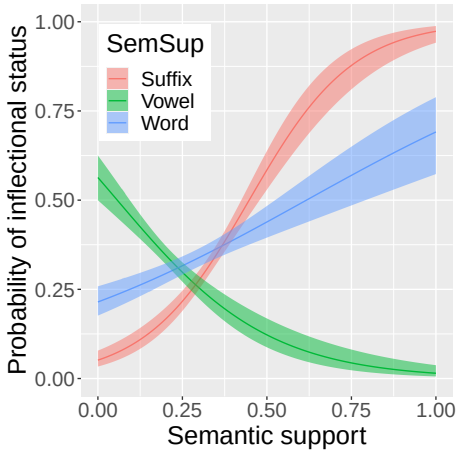
(a) Distributions of `SemSupSuffix` for inflected and non-inflected words.



(b) Distributions of `SemSupWord` for inflected and non-inflected words.



(c) Distributions of `SemSupVowel` for inflected and non-inflected words.



(d) Probability of inflectional status predicted by the three semantic support measures.

Figure 4.3: Comparison of `SemSupSuffix`, `SemSupVowel`, and `SemSupWord`.

support measures were always highly significant ($p < 0.001$).

As illustrated in Figure 4.3d, the effects of the three semantic support measures were qualitatively different. `SemSupSuffix` was associated the most strongly with the probability of inflectedness. The greater `SemSupSuffix` becomes, the more likely the word in question is to be inflected. A similar effect was observed also for `SemSupWord`, albeit to lesser degree. In contrast, higher `SemSupVowel` was correlated with lower probability of inflectedness.

In line with the results of the variable importances obtained with a Random Forest analysis above, the present analyses confirmed that SemSupSuffix was the most effective predictor for inflectional status. Accordingly, in the next section, we focus on SemSupSuffix to clarify whether SemSupSuffix is also predictive for tongue tip trajectories. Considering that SemSupSuffix was greater for inflected words and that inflected words showed articulatory enhancement (Section 4.2), greater SemSupSuffix is expected to be associated with articulatory enhancement. This hypothesis will be tested in the next section. In addition, the performance of SemSupSuffix is compared with that of the binary predictor inflectional status.

4.3.3 Predicting tongue trajectories from semantics

We used the same dataset as in Section 4.2 to compare the performance of semantic support for suffix (i.e., SemSupSuffix) with that of inflectional status as a binary predictor. Some words were not available in CELEX or the pre-trained word2vec model. As a consequence, 5.33% of the data points were lost.

For the remaining data, a GAMM was fitted with the same model structure as in section 4.2 except for the predictor for inflectional status. Inflectional status was represented by a binary factor in section 4.2 in interaction with normalized time and log-transformed frequency. In the following analyses, the binary factor was replaced with SemSupSuffix. We fitted the following model to the data, again including the three-way interaction:

```
TonguePosition ~ s(Time, Speaker, bs='fs', k=3, m=1) +
                 s(PrevSeg, bs='re', k=3) +
                 s(NextSeg, bs='re', k=3) +
                 s(VowelDuration, k=3) +
                 ti(VowelDuration, Time, k=c(3,3)) +
                 te(Time, SemSupSuffix, Freq, k=c(3, 3, 3))
```

The model with SemSupSuffix required one less edf, and nevertheless improved

the model fit significantly by 142.30 AIC units (by 62.73 ML scores), compared to the model with a binary factor of inflectional status[1]. All terms in the model were well-supported (except for the intercept; see Table 4.1). Figure 4.4 illustrates the interaction of `SemSupSuffix` by frequency at the center of the vowel. In this figure, the x axis represents frequency, and the y axis `SemSupSuffix`. Warmer colors represent higher tongue tip positions. Since the target vowel is [a(ː)], higher tongue positions correspond to articulatory reduction.

Table 4.1: Summary of the model with `SemSupSuffix`.

A. Parametric terms	Estimate	Std.Error	t-value	p-value
Intercept	2.628	2.530	1.038	0.299
B. Smooth terms	**edf**	**Ref.df**	**F**	**p-value**
s(Time, Speaker)	97.841	104.000	604.502	<0.001
s(PrevSeg)	16.671	17.000	547.183	<0.001
s(NextSeg)	7.870	8.000	2189.755	<0.001
s(VowelDuration)	1.005	1.010	18.850	<0.001
ti(Time, VowelDuration)	3.632	3.910	29.419	<0.001
te(Time, SemSupSuffix, Freq)	20.312	22.235	30.919	<0.001

Figure 4.4 shows that higher `SemSupSuffix` mainly goes hand in hand with lower tongue trajectories, indicating that the enhancement effect of `SemSupSuffix` is limited to higher frequency words. From the perspective of frequency effects, higher frequency is associated with higher tongue positions for low `SemSupSuffix` values, indicating articulatory reduction. In contrast, when `SemSupSuffix` is high, an increase in frequency is tied with lowering of the tongue tip, indicating articulatory enhancement. Since greater `SemSupSuffix` is correlated with inflectedness (Section 4.3.2), the current result is in line with the strong and attenuated reduction effects of frequency for non-inflected and inflected words respectively, reported in section 4.2.

Figure 4.4 does not show in details how tongue trajectories over time are mod-

[1]No model comparison test is necessary because the model with `SemSupSuffix` is simpler and better than the model with inflectional status as factor variable.

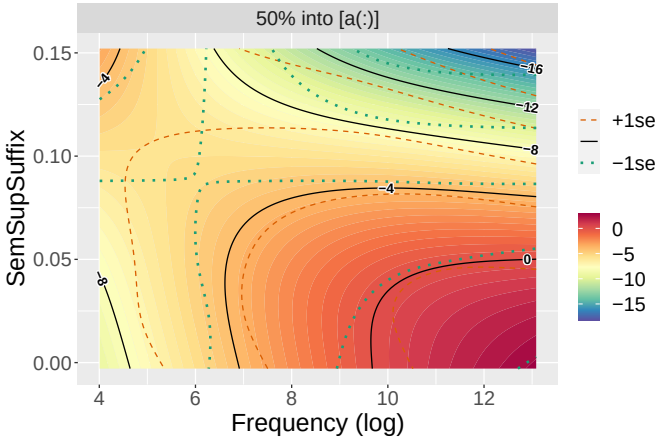


Figure 4.4: Tongue tip height as a function of frequency and `SemSupSuffix`. Time is fixed at 0.5 (at the middle of the vowel). Warmer colors represent high and colder colors represent low positions.

ulated by frequency and `SemSupSuffix`, because thus far time was fixed at the middle of the vowel. Figure 4.5 zooms in on time, illustrating qualitative differences in tongue trajectories as a function of time, frequency, and `SemSupSuffix`. For illustration, `SemSupSuffix` is discretized into high and low values, which correspond to 1% and 99% quantiles.

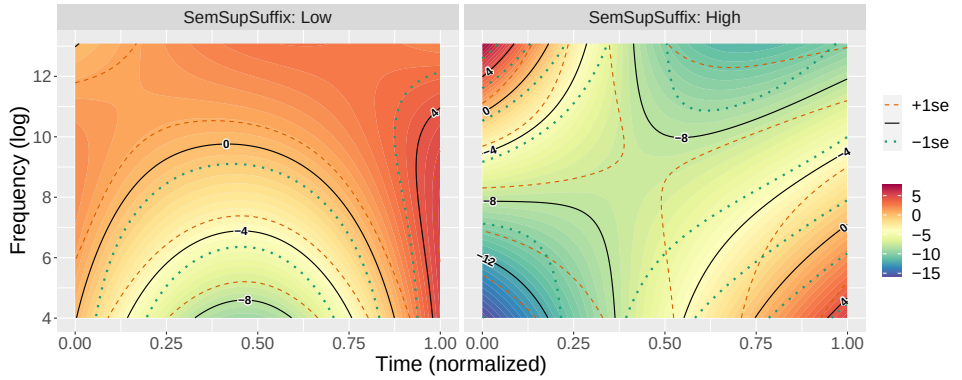


Figure 4.5: Tongue tip height as a function of time and frequency. `SemSupSuffix` is discretized to low and high values, which correspond to 0.01 and 0.99 quantiles. Warmer colors represent high and colder colors represents low positions.

When `SemSupSuffix` is low (in the left panel), lower frequency is associated with lower tongue trajectories, and higher frequency is associated with higher

tongue trajectories. On the other hand, when SemSupSuffix is high (in the right panel), high frequency words are articulated with tongue trajectories with a greater lowering from the middle to the onset of the vowel.

4.3.4 Interim summary

In this section, we showed that semantic support from the word to its word-final triphone (i.e. SemSupSuffix) outperformed other semantic measures in variable importance estimated by a Random Forest (Breiman, 2001). In line with this variable importance, SemSupSuffix also predicted inflectional status most effectively. Higher SemSupSuffix was associated with the word being inflected and also interacted with frequency. Because of the interaction with SemSupSuffix, higher frequency was associated with higher tongue trajectories, indicating reduced articulations, for low SemSupSuffix, while higher frequency was correlated with lower tongue positions, indicating enhanced articulations, for high SemSupSuffix. These observed patterns are in line with the patterns observed in the section 4.2. In 4.2, higher frequency was associated with strong articulatory reduction for non-inflected words, while the reduction effect was attenuated for inflected words.

Thus far, we have focused on the semantic support for the final triphone, which is centered around the exponent /-t/. We also considered a model in which the semantic support for the triphone straddling the vowel (i.e., SemSupVowel) is considered instead. This model shows that a greater semantic support for the vowel leads to a lower position of the tongue tip. At the same time, higher word frequency predicts higher tongue positions, irrespective of the amount of semantic support for the vowel (see Appendix 4.C for detail). As only 14 out of 70 word types have a stem that ends in a vowel, the vast majority of words have a vowel triphone that does not include the inflectional exponent. From these observations, we conclude that on the one hand, a greater semantic support for the vowel gives rise to enhanced articulation of the stem vowel, but that the effect of frequency works against this, giving rise to higher tongue positions.

4.4 Discussion

In what follows, we first explore possible explanations for the observed patterns. Subsequently, we propose our interpretation and lay out implications of our findings for existing theories.

Higher frequency has been reported to be correlated with both phonetic reduction (e.g. Aylett & Turk, 2004) as well as enhancement (e.g. Kuperman et al., 2007). These seemingly contradictory effects may be due to morphological status of the items being investigated. When the reduction effect is observed, morphologically simple words are always included. On the other hand, the enhancement effect has been observed only for morphologically complex words.

In order to clarify the role of morphological structure, we focused on inflected and non-inflected words in German. The target words shared the same rhyme structure, their stem vowel was [a(ː)], and their final segment was [t]. The word-final [t] was a part of the stem for non-inflected words, while it was the exponent for inflected words. Vertical tongue tip and body positions were fitted with Generalized Additive Mixed-effects Models (GAMMs) (Wood, 2017) as a function of time, frequency, and inflectional status together with random effect factors and control covariates.

The tongue tip and body models both showed significant effects of inflectional status. Inflected words showed lower tongue tip/body positions on average than non-inflected words. Since the vowel [a(ː)] was investigated and followed by a morpheme boundary in inflected words, these results suggest enhanced articulatory realizations in the pre-morpheme-boundary condition.

Pre-morpheme boundary enhancement is in harmony with the paradigm uniformity hypothesis (Seyfarth et al., 2017), which predicts that members of the same paradigm become similar in phonetic realizations to each other. However, the quality and degree of articulatory enhancement was significantly modulated by frequency. Inflected words retained lower tongue positions, namely more enhanced tongue positions, compared to non-inflected words, as frequency increased. This

interaction of inflectional status and frequency was observed for the tongue tip and the tongue body both.

Increased degrees of articulatory enhancement (implying decreased degrees of articulatory reduction) for higher frequency inflected words are consistent with the articulatory improvement hypothesis (Tomaschek, Tucker, et al., 2018). Higher frequency words are articulatorily well-practiced and therefore their articulations are faster and more enhanced. However, under this hypothesis, not only inflected words but also non-inflected words should be enhanced with increasing frequency. This, however, is not the case in the present study.

In the present study, we observed that non-inflected words were realized with greater degrees of articulatory reduction as frequency increases. This reduction effect is in line with the smooth signal redundancy hypothesis (Aylett & Turk, 2004). Higher frequency can go hand in hand with higher redundancy and lower amounts of information (surprisal). According to Aylett and Turk (2004), this motivates articulatory reduction. However, the smooth signal redundancy hypothesis does not take into consideration the morphological status of the word in question. Consequently, the hypothesis predicts the same degree of phonetic reduction also for inflected words, which was not the case in the present study.

Why do inflected words show less degrees of reduction, while non-inflected words show a strong reduction effect? One systematic difference between inflected and non-inflected words is the presence and absence of inflectional meanings. In German, inflectional meanings are mostly expressed by and tied in with inflectional suffixes. Strong form-meaning relations have been found to be a source of phonetic enhancement: (Gahl & Baayen, 2022) report that semantically better-supported words are realized with longer durations. This suggests that inflectional meanings may provide good semantic support for their corresponding inflectional suffixes, which in turn may lead to enhanced realizations in the corresponding inflected words.

This hypothesis was addressed, using the discriminative lexicon model (DLM:

Baayen et al., 2018; Baayen et al., 2019). Computational modeling revealed that the semantic support for the word-final triphone (SemSupSuffix) outperformed other semantic measures such as semantic support to the stem triphone. Greater SemSupSuffix was strongly associated with higher probability of inflectedness. Therefore, SemSupSuffix can be understood as a continuous counterpart of a categorical factor specifying inflectional status. Replacing the categorical predictor 'inflectional status' by SemSupSuffix resulted in a significant improvement in model fit.

SemSupSuffix was also shown to be predictive for vertical positions of the tongue tip. For higher-frequency words, a higher SemSupSuffix predicted a lower tongue position. From the perspective of the word frequency effect, SemSupSuffix emerges as a modulation of the word frequency effect. When SemSupSuffix was high, high frequency words were articulated with lower tongue positions. When SemSupSuffix was low, high frequency words were articulated with higher tongue positions. Since high SemSupSuffix was associated with inflected words, the modulation by SemSupSuffix explains why inflected words were less reduced, while non-inflected words showed strong reduction.

Importantly, this explanation does not require the theoretical concept of a 'morpheme boundary'. The present results therefore challenge the classical view of the speech production process such as formalized in the WEAVER++ model (Levelt et al., 1999; Levelt & Wheeldon, 1994; Roelofs, 1997), which operates on morphemes with at least one intermediate symbolic layer between semantics and phonetics. On the other hand, the present results support the hypothesis that better mappings between inflectional meanings and forms (inflectional suffixes) go hand in hand with enhanced realizations (Gahl & Baayen, 2022).

In the DLM model, the support from a words' meaning for the final /-t/ is much stronger for inflected words, which we have shown to be due to the inflectional semantics that are realized by this exponent. However, the enhancement observed for greater semantic support was observed for the vowel. This strong enhancement

of the vowel is likely to be due to co-articulation between the stem vowel and the suffix. This possibility is also supported by greater degrees of modulation of tongue trajectories by semantic support and frequency for tongue tip positions than for tongue body positions (compare Figure 4.5 in section 4.3.3 and Figure 4.C.2 in 4.C). Since the present study investigated [a(ː)] followed by the dental exponent [t], it makes sense that the co-articulation with [a(ː)] was more prominent for the tongue tip than the tongue body.

We observed that a higher semantic support for the vowel triphone, namely `SemSupVowel`, predicts lower positions for the tongue tip. However, the vowel triphone does not include the inflectional exponent. Unlike the final triphone, the vowel triphone is not systematically connected with the inflectional semantics of the /-t/ exponent. This may explain why, in a model replacing the final triphone with the triphone of the vowel, greater word frequency predicted higher positions of the tongue tip. It is only for the final triphone, and its co-articulatory entanglement with the preceding vowel, that the practice effect of frequency is visible.

Enhancement from semantic support clearly is not the only factor that co-determines articulation. For the non-inflected words, greater frequency goes hand in hand with higher tongue positions, which fits well with the argument of Aylett and Turk (2004) that less informative words reduce. For the inflected words in our dataset, we observed attenuated degrees of the reduction effect. This is likely due to the reduction effect being counterbalanced by the articulatory strengthening induced by the inflectional semantics.

It is possible to explain the reduction effect of predictability in the framework of the discriminative lexicon model. The present study showed that higher `SemSupSuffix`, namely higher $\hat{C}_{i,s}$, was correlated with phonetic enhancement. On the other hand, greater amount of information is said to also go hand in hand with enhanced realizations. Therefore, the effect of informativity can be integrated as a parameter modifying the strength of a semantic vector. Denoting the amount of information of a word ω at a point in a discourse k by $h_{\omega,k}$, the composite effect

of informativity and semantic support can be expressed as $h_{\omega,k}\hat{C}_{i,s}$ (see also Gahl & Baayen, 2022).

In summary, the present study shows how the paradox of two seemingly contradictory frequency effects can be resolved. Frequency effects can show up as different degrees of phonetic reduction, depending on morphological status. For inflected words, what looks like an attenuated reduction effect (and even a clear enhancement effect for some previous studies) is actually a composite of a reduction effect due to lack of informativity (e.g., Aylett & Turk, 2004) and a strengthening effect that is determined by the amount of semantic support that a word's form receives. For inflected words, this amount of support, especially for a word's final triphone, is driven by a word's inflectional semantics. In other words, what would seem to be an effect at the level of word form — a morphological boundary effect — actually is driven by inflectional semantics.

Appendices

4.A Assignment of inflectional status

CELEX tag	Example	Present study
0	*jemand*	non-inflected
13SIA	*stand*	non-inflected
2PIE,rP	*fangt*	inflected
2SIE,3SIE,2PIE	*aufpasst*	inflected
2SIE,3SIE,2PIE,rP	*kratzt*	inflected
2SIE,3SIE,2PIE,rP,pA	*erfasst*	inflected
3SIE,2PIE	*ausmacht*	inflected
3SIE,2PIE,pA	*ausbezahlt*	inflected
3SIE,2PIE,rP	*macht*	inflected
3SIE,2PIE,rP,pA	*bezahlt*	inflected
nP,gP,dP,aP,nS,dS,aS	*Watt*	non-inflected
nS	*Kandidat*	non-inflected
nS,dS,aS	*Land*	non-inflected
nS,dS,aS,nP,gP,dP,aP	*England*	non-inflected
nS,gS,dS,aS	*Hand*	non-inflected
pA	*gemacht*	inflected
pA,3SIE,2PIE,rP	*bestrahlt*	inflected
X	*bald*	non-inflected

4.B `SemSupSuffix` model for tongue body positions

A GAMM with the same structure as for `SemSupSuffix` (see section 4.3.3) was fitted to vertical tongue body positions. The interaction among time, `SemSupSuffix`, and frequency was supported as shown in the last row of Table 4.B.1 below.

A visualization of the interaction between frequency and `SemSupSuffix` at the center of the vowel (Figure 4.B.1) indicates that their effects are minimal in most combinations of values of `SemSupSuffix` and frequency. Patterns of tongue body trajectories are comparable for low and high values of `SemSupSuffix`, while higher frequency is constantly associated with higher tongue body positions.

The current dataset consists of the words with the stem vowel [a(ː)] and the

Table 4.B.1: Summary of the model with `SemSupSuffix` for tongue body positions.

A. Parametric terms	Estimate	Std.Error	*t*-value	*p*-value
Intercept	8.440	1.546	5.460	<0.001
B. Smooth terms	**edf**	**Ref.df**	***F***	***p*-value**
s(Time, Speaker)	57.310	104.000	393.725	<0.001
s(PrevSeg)	15.883	17.000	108.292	<0.001
s(NextSeg)	6.823	8.000	347.235	<0.001
s(VowelDuration)	1.002	1.005	0.182	0.672
ti(Time, VowelDuration)	3.041	3.569	5.663	<0.001
te(Time, SemSupSuffix, Freq)	13.826	14.973	9.651	<0.001

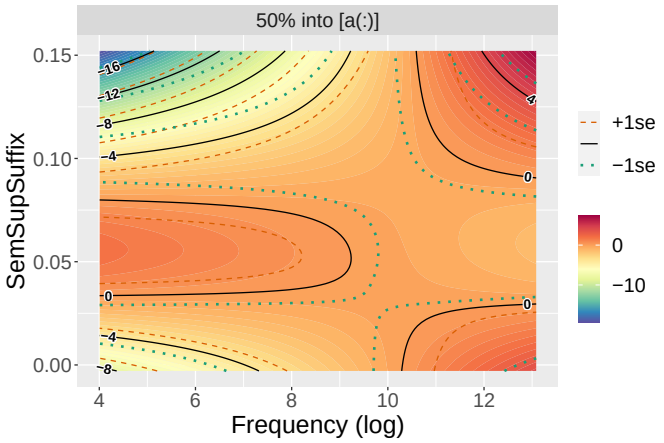


Figure 4.B.1: Tongue body height as a function of frequency and `SemSupSuffix`. Time is fixed at 0.5 (at the middle of the vowel). Warmer colors represent high and colder colors represent low positions.

word-final segment [t], which are expected to induce coarticulatory movements mainly for the tongue tip, leaving the tongue body being moved only passively. Therefore, these results suggest that strengthening effects by semantic support mainly influence coarticulatory movements of the tongue.

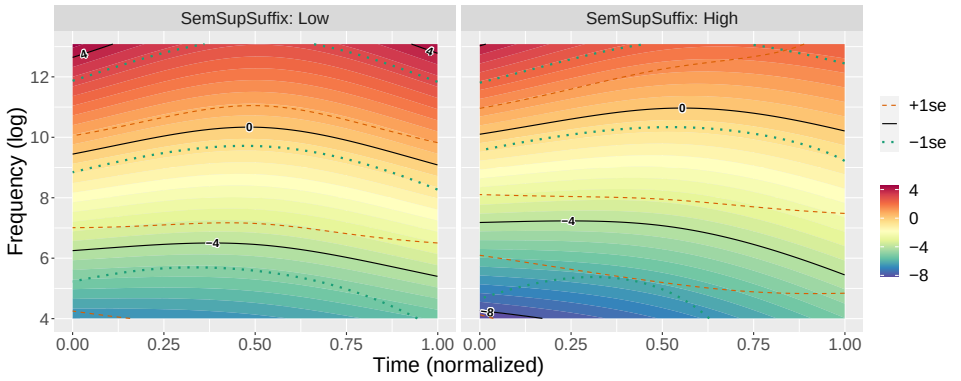


Figure 4.B.2: Tongue body height as a function of time and frequency. Semantic support for suffixes (i.e., `SemSupSuffix`) is discretized to low and high values, which correspond to 0.01 and 0.99 quantiles. Warmer colors represent high and colder colors represents low positions.

4.C `SemSupVowel` models

`SemSupVowel` was distributed in a right-skewed manner. Therefore, the variable was log-transformed in prior to fitting GAMMs. After the log-transformation, `SemSupVowel`was fitted with a GAMM to predict tongue tip and body positions with other control variables and random effects with the same model structure as for `SemSupSuffix` (see section 4.3.3 for the model structure), except for replacing `SemSupSuffix` for `SemSupVowel`.

4.C.1 Tongue tip

A fitted GAMM showed that higher `SemSupVowel` was constantly associated with lower tongue tip positions, namely clearer articulations (see Table 4.C.1 and Figure 4.C.1). Figure 4.C.2 further illustrates that higher frequency words are articulated with higher and flatter tongue tip trajectories when `SemSupVowel`is low, while higher frequency words show attenuated degrees of reduction (i.e., tongue-raising effects) when `SemSupVowel`is high.

Table 4.C.1: Summary of the model with log-transformed `SemSupVowel` for tongue tip positions.

A. Parametric terms	Estimate	Std.Error	*t*-value	*p*-value
Intercept	2.803	2.634	1.064	0.287
B. Smooth terms	**edf**	**Ref.df**	***F***	***p*-value**
s(Time, Speaker)	98.270	104.000	317.565	<0.001
s(PrevSeg)	13.718	14.000	507.679	<0.001
s(NextSeg)	7.801	8.000	1054.130	<0.001
s(VowelDuration)	1.793	1.955	12.690	<0.001
ti(Time, VowelDuration)	3.500	3.853	25.346	<0.001
te(Time, SemSupVowel, Freq)	24.158	25.261	34.194	<0.001

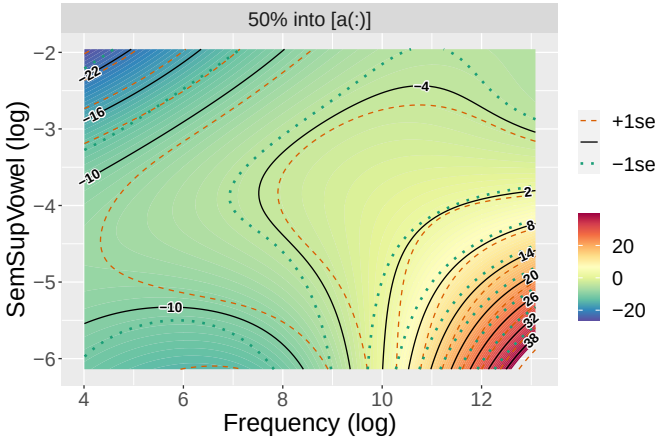


Figure 4.C.1: Tongue tip height as a function of frequency and log-transformed `SemSupVowel`. Time is fixed at 0.5 (at the middle of the vowel). Warmer colors represent high and colder colors represent low positions.

4.C.2 Tongue body

The same structure of a GAMM was fitted for tongue body positions (Table 4.C.2). The interaction of frequency and `SemSupVowel` turned out to be a U-shaped effect (Figure 4.C.3). This effect is likely due to extreme values predicted for very high and very low `SemSupVowel`values. For middle values of `SemSupVowel`, predicted tongue body height is almost always zero, indicating no substantial effect of frequency and `SemSupVowel`is visible in the region. In line with this observation, tongue trajectories are predicted to stay slightly higher than the occlusal plane (i.e.

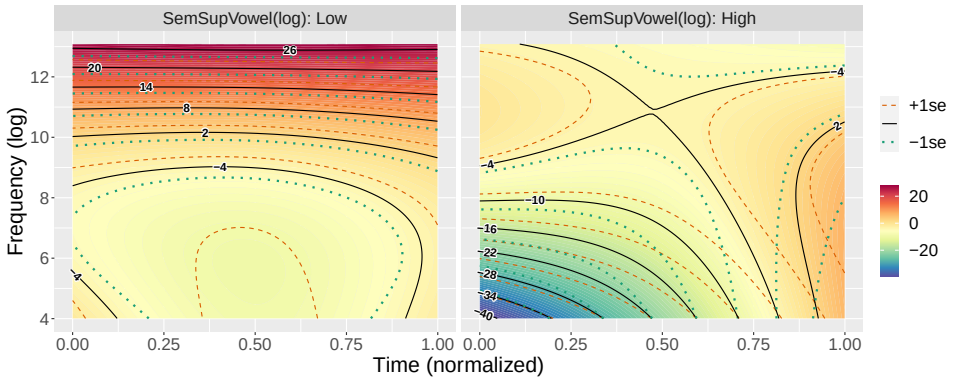


Figure 4.C.2: Tongue tip height as a function of time and frequency. `SemSupVowel` is log-transformed and discretized to low and high values, which correspond to 0.01 and 0.99 quantiles. Warmer colors represent high and colder colors represents low positions.

0) with not much raising or lowering during the vowel, regardless of values of frequency. A possible exception could be tongue body positions at the onset of the vowel for low frequency words with high `SemSupVowel`, where low positions are predicted. However, these predictions are not very reliable due to sparseness of data points below (log) frequency being 7.

Table 4.C.2: Summary of the model with log-transformed `SemSupVowel` for tongue body positions.

A. Parametric terms	Estimate	Std.Error	*t*-value	*p*-value
Intercept	8.312	1.401	5.935	<0.001
B. Smooth terms	**edf**	**Ref.df**	***F***	***p*-value**
s(Time, Speaker)	94.479	104.000	173.855	<0.001
s(PrevSeg)	11.962	14.000	44.459	<0.001
s(NextSeg)	6.076	8.000	108.862	<0.001
s(VowelDuration)	1.008	1.015	4.895	0.026
ti(Time, VowelDuration)	1.997	2.013	48.738	<0.001
te(Time, SemSupVowel, Freq)	23.823	25.147	25.439	<0.001

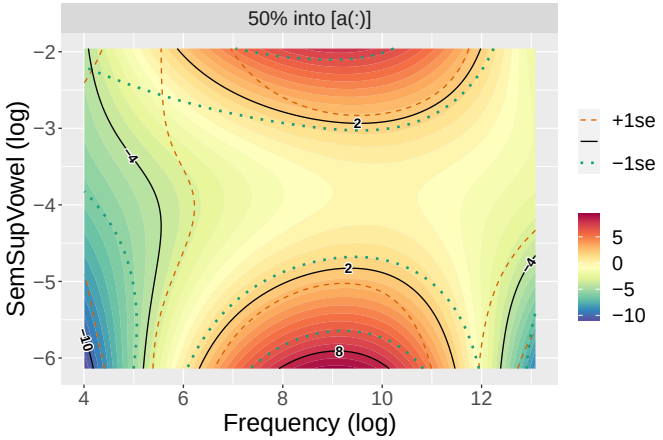


Figure 4.C.3: Tongue body height as a function of frequency and log-transformed `SemSupVowel`. Time is fixed at 0.5 (at the middle of the vowel). Warmer colors represent high and colder colors represent low positions.

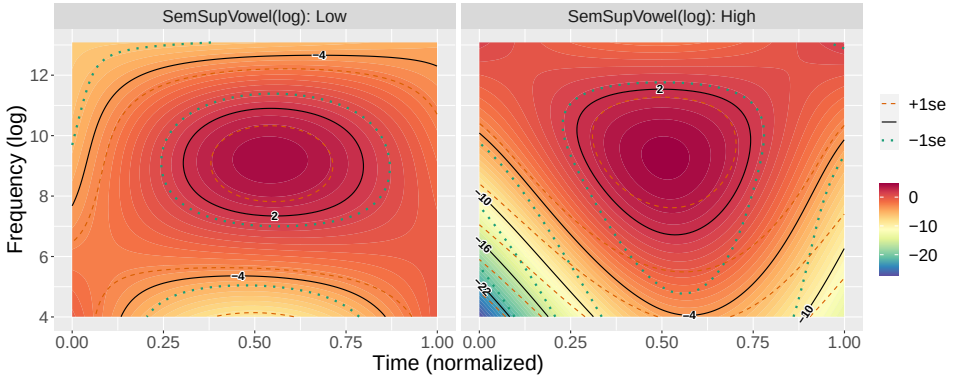


Figure 4.C.4: Tongue body height as a function of time and frequency. Semantic support for the stem vowel (i.e., `SemSupVowel`) is log-transformed and discretized to low and high values, which correspond to 0.01 and 0.99 quantiles. Warmer colors represent high and colder colors represents low positions.

Chapter 5

Summary and conclusions

Frequency has been reported to be associated with phonetically enhanced realizations such as longer duration and more peripheral tongue positions in a small number of studies (Cohen, 2014; Kuperman et al., 2007; Tomaschek, Tucker, et al., 2018; Tomaschek et al., 2021), whereas the large majority of studies report frequency to go hand in hand with phonetic reduction (Aylett & Turk, 2004, 2006; A. Bell et al., 2002; Dinkin, 2008; Gahl, 2008; Jurafsky et al., 2001; Lin et al., 2011; Pluymaekers et al., 2005b).

In this thesis, the finding that a higher frequency predicts phonetic enhancement was replicated using ultrasound, following up on one of the previous studies that found such an enhancement effect using electromagnetic articulography (Tomaschek, Tucker, et al., 2018). Consistent with the findings of this study, clearer articulations were observed for the [aː] vowel in high frequency words, compared to middle frequency words. Effects of frequency were more visible for the tongue tip than the tongue body. In the current study, as in the study of Tomaschek, Tucker, et al. (2018), the [aː] vowels were followed by alveolar suffixes, which explains why coarticulation was most prominently present for the tongue tip. The tongue body executed relatively passive movements that followed the relatively more active movements of the tongue tip.

In addition, effects of frequency were much more visible for the suffix condi-

tion [t], compared to [n]. This was most likely caused by possible differences in syllable structure. The suffix [n] condition contained [ən] and [n]. The former can be a syllabic nasal and the latter can be a separate syllable with a schwa. In either case, co-articulatory effects from the suffix onto the stem vowel are expected to become smaller. Considered jointly, this study clarifies that how strongly experience, as gauged with frequency, strengthens articulation can vary substantially with morphological context and the nature of the co-articulatory processes expected for these contexts.

The second study, in contrast, controlled for word-final syllable structure, and focused on the suffix [t] condition. Furthermore, in this study, words with non-morphemic word-final [t] were included in order to clarify whether frequency interacts with morphological structure. An interaction of frequency by morphological status (inflected vs. not inflected) indeed emerged from this study, which made use of electromagnetic articulography recordings of spontaneous conversational German. For non-inflected words, a higher frequency predicted greater phonetic reduction for the stem vowel (always [aː]). In contrast, when the final segment ([t]) was an inflectional exponent, the phonetic reduction effect was attenuated or even somewhat enhanced.

These observations support the possibility that presence or absence of a morphological boundary modulates frequency effects. In fact, these observations are consistent with the fact that in the literature, the words that show phonetic reduction for increasing frequency tend to be morphologically simple, whereas phonetic enhancement has only been reported for morphologically complex words.

However, the question remains: why do different morphological configurations create such different frequency effects? Neither an explanation based on morphological parsing with pressure from paradigm members, namely the paradigm uniformity account (Seyfarth et al., 2017), nor an explanation based on syntagmatic predictability (Aylett & Turk, 2004; Jurafsky et al., 2001) can account for the observed interaction of frequency by morphological status. For a better explanation,

we revisited the concept of "morphological boundary". Rather than assuming this construct as a theoretical primitive, we explored the possibility that this construct is itself grounded in semantics.

To this end, the third study investigated the relation between inflectional status and the amount of semantic support that an inflectional suffix receives from its meaning, using the DLM model. Compared to non-inflected words, inflected words received much greater semantic support for the word-final trigram, which straddled the inflectional exponent. Inflectional status was associated with greater semantic support for suffix. This finding is even more remarkable given that word2vec embeddings were used to represent words' meanings, rather than fastText vectors (which can 'look inside' word forms). Apparently, empirical word embeddings of inflected words ending in [t] co-vary systematically with the presence of this exponent in words' forms, allowing the DLM mapping from embeddings to trigrams to provide especially strong support for the trigram covering the inflectional exponent.

In addition, the third study of this dissertation revealed that greater semantic support for the suffixal trigram predicted phonetically more enhanced realization of the [aː]. Importantly, a regression model with semantic support as predictor provided a better fit to the data than a regression model with a factor specifying the absence or presence of an inflectional word boundary. This suggests that indeed semantics is the crucial factor at play, rather than a purely form-based 'invisible' boundary between stem and exponent.

Furthermore, an interaction of the semantic support measure by frequency was present, indicating that when semantic support for the final trigram was low (typical for non-inflected words) the vowel was more reduced, whereas when semantic support for the final trigram was high (typical for inflected words), the vowel was articulatorily enhanced. In other words, the measure of semantic support provided an explanation to the different degrees of phonetic reduction effects of frequency according to inflectional status observed in the second study without requiring the-

oretical constructs such as stems, exponents, and morphological boundaries. Such constructs are useful for high-level analyses, but for understanding the fine details of articulation, the distributional properties of words forms, their meanings, and the relation between these are crucial.

These findings cannot be explained by classical speech process models that assume serial processing based on modules such as syntax, morphology, phonology, and so on (Fromkin, 1971; Garrett, 1984, 1988; Levelt et al., 1999; Levelt & Wheeldon, 1994). In these models, morphological information, let alone semantics, is not available at the levels at which the details of phonetic realization are calculated. For some models, such as the Levelt model, it is a design feature that phonetic realization is completely shielded from semantics and syntax. Other models, such as the one proposed by Oppenheim et al. (2010), may be able to accommodate the present findings, but in order to do so with sufficient precision, it will be necessary to move from hand-crafted featural representations for word meaning to the embeddings of distributional semantics.

What I have shown in this doctoral dissertation is that a detailed quantitative assessment of how distributed semantics support low-level sublexical features of form can improve our understanding of the details of how we articulate words, without having to posit hierarchies of units mediating between meaning and form. Phoneticians tend to predict form from form. Semanticists tend to predict meaning from meaning. But language is a communication system that bridges the gap between form and meaning, and meaning and form. Current distributional semantics, in combination with the simple but highly effective algorithms of the DLM model, make it possible to understand in much greater detail how language bridges this gap. With current methods from statistics and machine learning, bridging this gap may be simpler than suggested by many models that thus far have been proposed as blueprints of human speech production.

Zusammenfassung und Fazit

In mehreren Studien wurde festgestellt, dass die Häufigkeit der Wörter in der Sprache mit phonetisch verstärkten Realisierungen verbunden sind. Beispielsweise haben häufige Wörter eine längere Dauer und eine peripherere Zungenpositionen (Cohen, 2014; Kuperman et al., 2007; Tomaschek, Tucker, et al., 2018; Tomaschek et al., 2021). Dies steht scheinbar im Widerspruch zu der vorherrschenden Ansicht, dass häufige Wörter phonetisch reduziert sind, das heißt, eine zentralisiertere Zungenpositionen besitzen und eine kürzere Dauer haben (Aylett & Turk, 2004, 2006; A. Bell et al., 2002; Dinkin, 2008; Gahl, 2008; Jurafsky et al., 2001; Lin et al., 2011; Pluymaekers et al., 2005b).

In dieser Doktorarbeit wurde zuerst der Verstärkungseffekt mittels Ultraschall reproduziert, der auch von einer der vorherigen Studie mit Elektromagnetische Artikulographie gefunden wurde (Tomaschek, Tucker, et al., 2018). In Übereinstimmung mit der Studie von Tomaschek, Tucker, et al. (2018) wurde in der ersten Studie dieser Doktroarbeit klarere Artikulationen für sehr häufige Wörter im Vergleich mit mittel häufigen Wörtern beobachtet. Die Effekte von der Worthäufigkeit waren sichtbarer für die Zungenspitze als für den Zungenkörper. In der aktuellen Studie, genauso wie in der Studie von Tomaschek, Tucker, et al. (2018), folgten dem [aː] Vokal Alveolarsuffixe, was erklärt, warum die Koartikulation an der Zungenspitze am stärksten ausgeprägt war. Der Zungenkörper führte nur nach der relativ aktiveren Bewegung der Zungenspitze relativ passive Bewegungen aus.

Außerdem waren Effekte der Worthäufigkeit in der aktuellen Studie viel sichtbarer für die Suffixbedingung [t] im Vergleich zu [n]. Dieser Unterschid wurde

höchstwahrscheinlich durch mögliche Unterschiede in den Silbenstrukturen verursacht. Die Suffixbedingung [n] enthielt [ən] und [n]. Ersteres kann ein Silbennasal und letzteres kann eine getrennte Silbe mit einem Schwa sein. In beiden Fällen sollten koartikulatorische Effekte vom Suffix auf den Stammvokal kleiner. Zusammengenommen verdeutlicht diese Studie, dass die Art und Weise, wie die Erfahrung, gemessen an der Worthäufigkeit, Artikulationen verstärkt. Allerdings kann die erwartete Koartikulation abhängig vom morphologischen Kontext erheblich unterschiedlich ausfallen.

Im Gegensatz kontrollierte die zweite Studie in dieser Doktorarbeit wortausschließende Silbenstrukturen, indem sich auf Suffixbedingung [t] konzentriert wurde. Darüber hinaus wurde in dieser Studie das nicht-morphämische wordausschließende [t] eingeschlossen, um die Wechselwirkung von Worthäufigkeit und morphologischem Status zu beobachten. Eine Wechselwirkung von Worthäufigkeit und morphologischem Status (flektiert vs. nicht flektiert) hat sich tatsächlich aus dieser Studie ergegeben, in der die Aufnahmen von Elektromagnetische Artikulographie von spontanem Konversationsdeutch benutzt wurden. Für nicht-flektierte Wörter sagte die höhere Frequenz stärkere phonetische Reduzierungen von dem Stammvokal (immer [aː]) vorher. Wenn das wortausschließende Segment ([t]) dagegen ein Flektionsexponent war, wurde der phonetische Reduzierungseffekt abgeschwächt und sogar etwas verstärkt.

Diese Beobachtungen unterstützen die Möglichkeit, dass das Vorhandensein/Fehlen von einer morphologischen Grenze Effekte von Worthäufigkeit moduliert. Tatsächlich stimmen diese Beobachtungen damit überein, dass in der Literatur die Wörter, die den Reduktionseffekt für höhere Frequenzen zeigten, tendenziell morphologisch einfach sind, während der phonetische Verstärkungseffekt nur für morphologisch komplexe Wörter gefunden wurde.

Aber warum verursachen unterschiedliche morphologische Strukturen solche unterschiedlichen Effekte in Abhängigkeit der Worthäufigkeit? Weder die Erklärung, die auf morphologischem Parsing (Syntaxanalyse) mit Druck von Paradig-

menmitgliedern basiert, die Paradigm-Uniformity Hypothese, noch die Erklärung, die auf syntagmatischer Vorhersagbarkeit basiert (Aylett & Turk, 2004; Jurafsky et al., 2001), liefern ausreichende Erkärungen für die beobachte Wechelwirkung von Worthäufigkeit und morphologischem Status. Zur besseren Erklärung haben wir den Begriff der *morphologischen Grenze* übergedacht. Dafür haben wir die morphologische Grenze nicht als gegeben angenommen, sondern sind davon ausgegangen, dass die morphologische Grenze durch eine semantische Ebene markiert wird.

Um zu überprüfen, ob die morphologische Grenze durch einen semantischen Einfluss vermittelt sein kann, untersuchte die dritte Studie in dieser Doktorarbeit mithilfe des DLM-Modells die Beziehung zwischen dem Flexionsstatus und der Höhe der semantischen Unterstützung, die das Suffix durch die Bedeutung des Worts erhält. Im Vergleich mit nicht-flektierten Wörtern erhielten flektierte Wörter deutlich mehr semantische Unterstützung für das wortausschließende Trigramm, das dem Flektionsexponent überspannte. Das Flektionsstatus war mit der höheren semantischen Unterstützung verbunden. Dieses Ergebnis ist umso bemerkenswerter, wenn man denkt, dass Einbettungen nicht von fastText, die in Wortformen 'hineinschauen' können, sondern von word2vec eingerichtet wurden, um Bedeutungen von Wörtern darzustellen. Offensichtlich variieren empirische Worteinbettungen von flektieren Wörtern, die mit [t] enden, systematisch mit dem Vorhandensein von diesem Flektionsexponent in Wortformen, was es dem DLM-Modell ermöglicht, welches Einbettungen auf Trigramme abbildet, besonders starke Unterstützungen für Trigramme bereitzustellen, die Flektionsexponenten abdecken.

Zusätzlich ergab die dritte Studie dieser Doktorarbeit, dass höhere semantische Unterstützung für das Suffixtrigramm phonetisch stärkere Realisierungen von dem Stammvokal [aː] vorhersagte. Wichtig ist, dass das Regressionsmodell mit semantischer Unterstützung als Prädiktor eine bessere Anpassung an die Daten lieferte, als das Regressionsmodell mit einem Faktor, der das Vorhandensein und Fehlen

von einer Flektionswortgrenze angibt. Diese Ergebnisse deuten darauf hin, dass tatsächlich Semantik der entscheidende Faktor ist und nicht eine rein formbasierte 'unsichtbare' Grenze zwischen Stamm und Exponent.

Darüber hinaus wurde die Wechselwirkung der semantischen Unterstützung und Worthäufigkeit beobachtet, was darauf hindeutet, dass der Stammvokal mehr reduziert war, wenn die semantische Unterstützung für das wortausschließende Trigramm niedrig war (typisch für nicht-flektierte Wörter), während der Stammvokal artikulatorisch verstärkt wurde, wenn die semantische Unterstützung für das wortausschließende Trigramm hoch war (typisch für flektierte Wörter). Mit anderen Worten lieferte das Maß von der semantischen Unterstützung die Erklärung für die nach dem Flektionsstatus unterschiedlichen Grade der phonetischen Reduktionseffekte der Worthäufigkeit, die in der zweiten Studie in dieser Doktorarbeit beobachtet wurden, ohne dass theoretische Begriffe wie Stämme, Exponenten, und morphologische Grenzen erforderlich sind. Solche Begriffe sind für Analysen auf hoher kognitiver Verarbeitungsebene nützlich. Aber um feine Details von Artikulationen zu verstehen, sind Verteilungseigenschaften von Wortformen, ihre Bedeutungen, und das Zusammenhang zwischen diesen entscheidend.

Diese neuen Erkenntnisse lassen sich nicht einfach durch klassische Sprachprozessmodelle erklären, die von der auf Modulen wie Syntax, Morphologie, Phonologie, usw. basierten seriellen Verarbeitungssytem ausgehen (Fromkin, 1971; Garrett, 1984, 1988; Levelt et al., 1999; Levelt & Wheeldon, 1994). In diesen Modellen stehen morphologische Informationen, geschweige denn semantische Informationen, nicht zur Verfügung, um feine phonetische Details zu bestimmen. Für einige Modelle, beispielsweise das Level Modell, ist es ein Designmerkmal, dass die phonetische Realisierung vollständig von Semantik und Syntax abgeschirmt ist. Andere Modelle, beispielsweise das Modell von Oppenheim et al., 2010, können möglicherweise die vorliegenden Ergebnisse berücksichtigen. Aber um das mit ausreichender Präzision zu erreichen würde es nötig sein, von handgefertigen Merkmalsdarstellungen für Wortbedeutungen zu Worteinbettungen

von Verteilungssemantik überzugehen.

Mit dieser Doktorarbeit habe ich gezeigt, dass eine detaillierte quantitative Bewertung von der Art und Weise, wie Verteilungssemantik sublexikalische Merkmale von Formen auf niedriger Ebene unterstützt, unser Verständnis darüber verbessern kann, wie man Wörter artikuliert, ohne Hierarchien von Einheiten zwischen Bedeutung und Form postulieren zu müssen. Phonetiker sagen — vereinfacht ausgedrückt — Form von Form vorher. Semantiker sagen — vereinfacht ausgedrückt — Bedeutung von Bedeutung vorher. Aber die Sprache ist ein Kommunikationssystem, das die Lücke zwischen Form und Bedeutung, sowie zwischen Bedeutung und Form, überbrückt. Die aktuelle Verteilungssemantik in Kombination mit den einfachen aber äußerst effektiven Algorithmen des DLM Modells ermöglicht es, besser zu verstehen, wie die Sprache diese Lücke überbrückt. Mithilfe aktueller Methoden aus Statistik und Machine-Learning ist die Überbrückung von dieser Lücke einfacher, als viele Modelle vermuten lassen, die bisher als Blaupausen der menschlichen Sprachproduktion vorgeschlagen wurden.

Bibliography

Arnold, D., & Tomaschek, F. (2016). The Karl Eberhards Corpus of spontaneously spoken southern German in dialogues — audio and articulatory recordings. *Tagungsband der 12. Tagung Phonetik und Phonologie im deutschsprachigen Raum*, 9–11.

Arnon, I., & Cohen Priva, U. (2013). More than words: The effect of multi-word frequency and constituency on phonetic duration. *Language and Speech*, *56*(3), 349–371. https://doi.org/10.1177/0023830913484891

Articulate Instruments Ltd. (2012). Articulate Assistant Advanced user guide: Version 2.14.

Aubin, J., & Ménard, L. (2006). Compensation for a labial perturbation: An acoustic and articulatory study of child and adult French speakers. In H. C. Yehia, D. Demolin, & R. Laboissiere (Eds.), *Proceedings of ISSP 2006: 7th International Seminar on Speech Production* (pp. 209–216).

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, *47*(1), 31–56. https://doi.org/10.1177/00238309040470010201

Aylett, M., & Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *The Journal of the Acoustical Society of America*, *119*(5), 3048–3058. https://doi.org/10.1121/1.2188331

Baayen, R. H., & Moscoso del Prado Martín, F. (2005). Semantic density and past-tense formation in three Germanic languages. *Language*, *81*, 666–698.

Baayen, R. H., Chuang, Y.-Y., & Blevins, J. P. (2018). Inflectional morphology with linear mappings. *The Mental Lexicon*, *13*(2), 230–268. https://doi.org/10.1075/ml.18010.baa

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., & Blevins, J. P. (2019). The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning. *Complexity*, 1–39. https://doi.org/10.1155/2019/4895891

Baayen, R. H., Dijkstra, T., & Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language*, *37*(1), 94–117. https://doi.org/10.1006/jmla.1997.2509

Baayen, R. H., Feldman, L. B., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, *55*(2), 290–313. https://doi.org/10.1016/j.jml.2006.03.008

Baayen, R. H., Kuperman, V., & Bertram, R. (2010). Frequency effects in compound processing. In S. Scalise & I. Vogel (Eds.), *Cross-disciplinary issues in compounding* (pp. 257–270). John Benjamins.

Baayen, R. H., & Linke, M. (2020). An introduction to the generalized additive model. In M. Paquot & S. T. Gries (Eds.), *A practical handbook of corpus linguistics* (pp. 563–591). Springer.

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*(3), 438–481. https://doi.org/10.1037/a0023851

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

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical database (release 2). *Published by the Linguistic Data Consortium, University of Pennsylvania*.

Baayen, R. H., & Smolka, E. (2020). Modeling morphological priming in German with naive discriminative learning. *Frontiers in Communication*, *5*(April). https://doi.org/10.3389/fcomm.2020.00017

Baayen, R. H., Tweedie, F. J., & Schreuder, R. (2002). The subjects as a simple random effect fallacy: Subject variability and morphological family effects in the mental lexicon. *Brain and Language*, *81*(1-3), 55–65. https://doi.org/10.1006/brln.2001.2506

Baayen, R. H., Vasishth, S., Kliegl, R., & Bates, D. (2017). The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, *94*, 206–234. https://doi.org/10.1016/j.jml.2016.11.006

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*(1), 92–111. https://doi.org/10.1016/j.jml.2008.06.003

Bell, A., Gregory, M. L., Brenier, J. M., Jurafsky, D., Ikeno, A., & Girand, C. (2002). Which predictability measures affect content word durations? *Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology (PMLA), ISCA Tutorial and Research Workshop (ITRW)*, 1–5.

Bell, M. J., Ben Hedia, S., & Plag, I. (2021). How morphological structure affects phonetic realisation in English compound nouns. *Morphology*, *31*(2), 87–120. https://doi.org/10.1007/s11525-020-09346-6

Bertram, R., Schreuder, R., & Baayen, R. H. (2000). The balance of storage and computation in morphological processing: The role of word formation type, affixal homonymy, and productivity. *Journal of Experimental Psychology: Learning Memory and Cognition*, *26*(2), 489–511. https://doi.org/10.1037/0278-7393.26.2.489

Blevins, J. P. (2016). *Word and paradigm morphology*. Oxford University Press.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Botchu, R., Bharath, A., Davies, A. M., Butt, S., & James, S. L. (2018). Current concept in upright spinal MRI. *European Spine Journal*, *27*(5), 987–993. https://doi.org/10.1007/s00586-017-5304-3

Brandt, E., Möbius, B., & Andreeva, B. (2021). Dynamic formant trajectories in German read speech: Impact of predictability and prominence. *Frontiers in Communication*, *6*, 1–15. https://doi.org/10.3389/fcomm.2021.643528

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Bressmann, T., Thind, P., Uy, C., Bollig, C., Gilbert, R. W., & Irish, J. C. (2005). Quantitative three-dimensional ultrasound analysis of tongue protrusion, grooving and symmetry: Data from 12 normal speakers and a partial glossectomee. *Clinical Linguistics and Phonetics*, *19*(6-7), 573–588. https://doi.org/10.1080/02699200500113947

Buchaillard, S., Perrier, P., & Payan, Y. (2009). A biomechanical model of cardinal vowel production: Muscle activations and the impact of gravity on tongue positioning. *The Journal of the Acoustical Society of America*, *126*(4), 2033. https://doi.org/10.1121/1.3204306

Carra, B. J., Bui-Mansfield, L. T., O'Brien, S. D., & Chen, D. C. (2014). Sonography of musculoskeletal soft-tissue masses: Techniques, pearls, and pitfalls. *American Journal of Roentgenology*, *202*(6), 1281–1290. https://doi.org/10.2214/AJR.13.11564

Cho, T. (2001). Effects of morpheme boundaries on intergestural timing: Evidence from Korean. *Phonetica*, *58*, 129–162.

Chomsky, N. (1965). *Aspects of the theory of syntax*. The MIT press.

Chomsky, N. (1981). *Lectures on government and binding: The pisa lectures*. Mouton de Gruyter.

Chuang, Y.-Y., & Baayen, R. H. (2021). Discriminative learning and the lexicon: NDL and LDL. In *Oxford research encyclopedia of linguistics*. Oxford University Press.

Chuang, Y.-Y., Kang, M., Luo, X., & Baayen, R. H. (2022). Vector space morphology with linear discriminative learning. In D. Crepaldi (Ed.), *Linguistic morphology in the mind and brain*. Routledge. http://arxiv.org/abs/2107.03950

Chuang, Y.-Y., Lõo, K., Blevins, J. P., & Baayen, R. H. (2020). Estonian case inflection made simple: A case study in Word and Paradigm morphology with linear discriminative learning. In L. Körtvélyessy & P. Štekauer (Eds.), *Complex words: Advances in morphology* (pp. 119–141). Cambridge University Press.

Chuang, Y.-Y., Vollmer, M. L., Shafaei-Bajestan, E., Gahl, S., Hendrix, P., & Baayen, R. H. (2021). The processing of pseudoword form and meaning in production and comprehension: A computational modeling approach using linear discriminative learning. *Behavior Research Methods*, *53*, 945–976. https://doi.org/10.3758/s13428-020-01356-w

Chuang, Y.-Y., Vollmer, M.-l., Shafaei-bajestan, E., Gahl, S., Hendrix, P., & Baayen, R. H. (2019). On the processing of nonwords in word naming and auditory lexical decision. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia* (pp. 1233–1237). Australasian Speech Science; Technology Association.

Cohen, C. (2014). Probabilistic reduction and probabilistic enhancement: Contextual and paradigmatic effects on morpheme pronunciation. *Morphology*, *24*(4), 291–323. https://doi.org/10.1007/s11525-014-9243-y

Cohen Priva, U. (2015). Informativity affects consonant duration and deletion rates. *Laboratory Phonology*, *6*(2), 243–278. https://doi.org/10.1515/lp-2015-0008

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428.

Cornell Statistical Consulting Unit. (2020). Overlapping confidence intervals and statistical significance, 2–4.

Dang, J., Lu, X., Tiede, M., & Honda, K. (2008). Inherent vowel structures in speech production and perception spaces. *Proceedings of ISSP 2008 - 8th International Seminar on Speech Production*, 37–40.

Dang, J., Tiede, M., & Yuan, J. (2009). Comparison of vowel structures of Japanese and English in articulatory and auditory spaces. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, (4), 2815–2818.

Davidson, L. (2005). Addressing phonological questions with ultrasound. *Clinical Linguistics and Phonetics*, *19*(6-7), 619–633.

Davidson, L. (2006). Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance. *The Journal of the Acoustical Society of America*, *120*(1), 407–415.

Dawson, K. M., Tiede, M. K., & Whalen, D. H. (2016). Methods for quantifying tongue shape and complexity using ultrasound imaging. *Clinical Linguistics and Phonetics*, *30*(3-5), 328–344. https://doi.org/10.3109/02699206.2015.1099164

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283–321. https://doi.org/10.1037//0033-295x.93.3.283

Dell, G. S. (1988). The retrieval of phonological forms in production: Tests of predictions from a connectionist model. *Journal of Memory and Language*, *27*(2), 124–142. https://doi.org/10.1016/0749-596X(88)90070-8

Dell, G. S. (1990). Effects of frequency and vocabulary type on phonological speech errors. *Language and Cognitive Processes*, *5*(4), 313–349. https://doi.org/10.1080/01690969008407066

Dell, G. S., Martin, N., & Schwartz, M. F. (2007). A case-series test of the interactive two-step model of lexical access: Predicting word repetition from picture naming. *Journal of Memory and Language*, *56*(4), 490–520. https://doi.org/10.1016/j.jml.2006.05.007

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review*, *104*(4), 801–838.

Denistia, K., & Baayen, R. H. (2022). The morphology of Indonesian: Data and quantitative modeling. In C. Shei & S. Li (Eds.), *The Routledge handbook of Asian linguistics* (pp. 605–634). Routledge.

Dinkin, A. J. (2008). The real effect of word frequency on phonetic variation. *Proceedings of the 31st Annual Penn Linguistics Colloquium*, *14*(1), 97–106.

Duñabeitia, J. A., Perea, M., & Carreiras, M. (2007). The role of the frequency of constituents in compound words: Evidence from Basque and Spanish. *Psychonomic Bulletin and Review*, *14*(6), 1171–1176. http://link.springer.com/10.3758/BF03193108

Epstein, M., Hacopian, N., & Ladefoged, P. (2002). Dissection of the speech production mechanism. *UCLA Working Papers in Phonetics*, *102*.

Erickson, D., Kawahara, S., Shibuya, Y., Suemitsu, A., & Tiede, M. (2014). Comparison of jaw displacement patterns of Japanese and American speakers of English: A preliminary report. *Journal of the Phonetic Society of Japan*, *18*(2), 88–94. https://ci.nii.ac.jp/naid/110009872251/

Faaß, G., & Eckart, K. (2013). SdeWaC: A corpus of parsable sentences from the web. In I. Gurevych, C. Biemann, & T. Zesch (Eds.), *Language processing and knowledge in the web* (pp. 61–68). Springer.

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior*, *12*(6), 627–635. https://doi.org/10.1016/S0022-5371(73)80042-8

Foygel, D., & Dell, G. S. (2000). Models of impaired lexical access in speech production. *Journal of Memory and Language*, *43*(2), 182–216. https://doi.org/10.1006/jmla.2000.2716

Fromkin, V. A. (1971). The non-anomalous nature of anomalous utterances. *Language*, *47*(1), 27–52. https://doi.org/10.2307/412187

Gahl, S. (2008). *Time* and *thyme* are not homophones: the effect of lemma frequency on word durations in spontaneous speech. *Language*, *84*(3), 474–496.

Gahl, S., & Baayen, R. H. (2022). Time and thyme again: Connecting spoken word duration to models of the mental lexicon. *OSF*, 1–41. https://osf.io/2bd3r/

Gahl, S., & Garnsey, S. M. (2004). Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language*, *80*(4), 748–775. https://doi.org/10.1353/lan.2004.0185

Gahl, S., Yao, Y., & Johnson, K. (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, *66*(4), 789–806. https://doi.org/10.1016/j.jml.2011.11.006

Gardner, M. K., Rothkopf, E. Z., Lapan, R., & Lafferty, T. (1987). The word frequency effect in lexical decision: Finding a frequency-based component. *Memory and Cognition*, *15*(1), 24–28.

Garrett, M. F. (1984). The organization of processing structure for language production: Applications to aphasic speech. In D. Caplan, A. R. Lecours, & A. Smith (Eds.), *Biological perspectives on language* (pp. 172–193). MIT press.

Garrett, M. F. (1988). Processes in language production. In F. J. Newmeyer (Ed.), *Linguistics: The Cambridge Survey Volume III: Language: Psychological and biological aspects* (pp. 69–96). Cambridge University Press.

Gick, B. (2002). The use of ultrasound for linguistic phonetic fieldwork. *Journal of the International Phonetic Association*, *32*(2), 113–121. https://doi.org/10.1017/S0025100302001007

Gordon, J. K. (2002). Phonological neighborhood effects in aphasic speech errors: Spontaneous and structured contexts. *Brain and Language*, *82*(2), 113–145.

Gu, C. (2002). *Smoothing Spline ANOVA Models*. Springer.

Harley, T. A., & Bown, H. E. (1998). What causes a tip-of-the-tongue state? Evidence for lexical neighbourhood effects in speech production. *British Journal of Psychology*, *89*, 151–174.

Harshman, R., Ladefoged, P., & Goldstein, L. (1977). Factor analysis of tongue shapes. *The Journal of the Acoustical Society of America*, *62*(3).

Hay, J. (2007). The phonetics of 'un'. In J. Munat (Ed.), *Lexical creativity, texts and contexts* (pp. 39–57). John Benjamins.

Hay, J. (2003). *Causes and consequences of word structure*. Routledge.

Hayes, B. P. (2000). Gradient well-formedness in optimality theory. In J. Dekkers, F. van der Leeuw, & J. van de Weijer (Eds.), *Optimality theory: Phonology, syntax, and acquisition* (pp. 88–120). Oxford University Press.

Heitmeier, M., & Baayen, R. H. (2020). Simulating phonological and semantic impairment of English tense inflection with Linear Discriminative Learning. *The Mental Lexicon*, *15.3*, 385–421.

Heitmeier, M., Chuang, Y., Axen, S., & Baayen, R. H. (2022). Frequency-informed linear discriminative learning. *Proceedings of a workshop in honor of Ingo Plag's 60-th birthday*.

Heitmeier, M., Chuang, Y.-Y., & Baayen, R. H. (2021). Modeling morphology with linear discriminative learning: Considerations and design choices. *Frontiers in Psychology*, *12*, 1–39.

Hertrich, I., & Ackermann, H. (2000). Lip–jaw and tongue–jaw coordination during rate-controlled syllable repetitions. *The Journal of the Acoustical Society of America*, *107*(4), 2236–2247. https://doi.org/10.1121/1.428504

Heyne, M., Derrick, D., & Al-Tamimi, J. (2019). Native language influence on brass instrument performance: An application of Generalized Additive Mixed Models (GAMMs) to midsagittal ultrasound images of the tongue. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02597

Hiiemae, K. M., Palmer, J. B., Medicis, S. W., Hegener, J., Scott Jackson, B., & Lieberman, D. E. (2002). Hyoid and tongue surface movements in speaking and eating. *Archives of Oral Biology*, *47*(1), 11–27. https://doi.org/10.1016/S0003-9969(01)00092-9

Hoedl, P. (2015). Defying gravity: Formant frequencies of English vowels produced in upright and supine body position. *Proceedings of the 18th International Congress of Phonetic Sciences*.

Howson, P. J., & Redford, M. A. (2019). Liquid coarticulation in child and adult speech. *Proceedings of the 19th International Congress of Phonetic Sciences*.

Jescheniak, J. D., & Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning Memory and Cognition*, *20*(4), 824–843.

Jordan, A. S., & White, D. P. (2008). Pharyngeal motor control and the pathogenesis of obstructive sleep apnea. *Respiratory Physiology and Neurobiology*, *160*(1), 1–7. https://doi.org/10.1016/j.resp.2007.07.009

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In

J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 229–254). John Benjamins.

Kelso, J. A. S., Vatikiotis-Bateson, E., Saltzman, E. L., & Kay, B. (1985). A qualitative dynamic analysis of reiterant speech production: Phase portraits, kinematics, and dynamic modeling. *Journal of the Acoustical Society of America*, *77*(1), 266–280. https://doi.org/10.1121/1.392268

Kim, A. M., Keenan, B. T., Jackson, N., Chan, E. L., Staley, B., Poptani, H., Torigian, D. A., Pack, A. I., & Schwab, R. J. (2014). Tongue fat and its relationship to obstructive sleep apnea. *Sleep*, *37*(10), 1639–1648D. https://doi.org/10.5665/sleep.4072

Kittredge, A. K., Dell, G. S., Verkuilen, J., & Schwartz, M. F. (2008). Where is the effect of frequency in word production? Insights from aphasic picture-naming errors. *Cognitive Neuropsychology*, *25*(4), 463–492. https://doi.org/10.1080/02643290701674851

Krott, A., Baayen, R. H., & Schreuder, R. (2001). Analogy in morphology: Modeling the choice of linking morphemes in Dutch. *Linguistics*, *39*(371), 51–93. https://doi.org/10.1515/ling.2001.008

Krott, A., Schreuder, R., Baayen, R. H., & Dressler, W. U. (2007). Analogical effects on linking elements in German compound words. *Language and Cognitive Processes*, *22*(1), 25–57.

Kuehn, D. P., & Moll, K. L. (1976). A cineradiographic study of VC and CV articulatory velocities. *Journal of Phonetics*, *4*(4), 303–320. https://doi.org/10.1016/s0095-4470(19)31257-4

Kuhn, M. (2021). caret: Classification and regression training. https://cran.r-project.org/package=caret

Kuperman, V., Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2007). Morphological predictability and acoustic duration of interfixes in Dutch compounds. *The Journal of the Acoustical Society of America*, *121*(4), 2261–2271. https://doi.org/10.1121/1.2537393

Kutzner, E. A., Miot, C., Liu, Y., Renk, E., Park, J. S., & Inman, J. C. (2017). Effect of genioglossus, geniohyoid, and digastric advancement on tongue base and hyoid position. *Laryngoscope*, *127*(8), 1938–1942. https://doi.org/10.1002/lary.26380

Lee, J., Kim, S., & Cho, T. (2019). Effects of morphological structure on intergestural timing in different prosodic-structural contexts in Korean. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*. Australasian Speech Science; Technology Association Inc.

Lee-Kim, S.-I., Davidson, L., & Hwang, S. (2013). Morphological effects on the darkness of English intervocalic /l/. *Laboratory Phonology*, *4*(2), 475–511. https://doi.org/10.1515/lp-2013-0015

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1–75.

Levelt, W. J. M., & Wheeldon, L. (1994). Do speakers have access to a mental syllabary? *Cognition*, *50*, 239–269.

Li, M., Kambhamettu, C., & Stone, M. (2005). Automatic contour tracking in ultrasound images. *Clinical Linguistics and Phonetics*, *19*(6-7), 545–554. https://doi.org/10.1080/02699200500113616

Li, V. G., Oh, S., Chopra, G., Celli, J., & Shaw, J. A. (2020). Articulatory correlates of morpheme boundaries: Preliminary evidence from intra- and intergestural timing in the articulation of the English past tense. *Proceedings of ISSP 2020 - 12th International Seminar on Speech Production*.

Liljencrants, J. (1971). A Fourier series description of the tongue profile. *Speech transmission laboratory - Quarterly progress and status report (STL-QPSR)*, *12*(4), 9–18.

Lin, S. S., Beddor, P. S., & Coetzee, A. W. (2011). Gestural reduction and sound change: An ultrasound study. *ICPhS XVII: Proceedings of the 17th International Congress of Phonetic Sciences*, 1250–1253.

Lindblom, B. (1983). Economy of speech gestures. In P. F. MacNeilage (Ed.), *The production of speech* (pp. 217–245). Springer-Verlag.

Lindblom, B., & Marchal, A. (1990). Explaining phonetic variation: A sketch of the H&H Theory. In W. J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modelling* (pp. 403–439). Kluwer Academic Publishers.

Lohmann, A. (2018a). *Cut* (n) and *cut* (v) are not homophones: lemma frequency affects the duration of noun–verb conversion pairs. *Journal of Linguistics*, *54*(4), 753–777. https://doi.org/10.1017/s0022226717000378

Lohmann, A. (2018b). Time and thyme are not homophones: A closer look at Gahl's work on the lemma-frequency effect, including a reanalysis. *Language*, *94*(2), e180–e190. https://doi.org/10.1353/lan.2018.0032

Luo, X., Chuang, Y.-Y., & Baayen, R. H. (2021). JudiLing: An implementation in Julia of Linear Discriminative Learning algorithms for language model.

Ma, J. K.-Y., & Wrench, A. A. (2022). Automated assessment of hyoid movement during normal swallow using ultrasound. *International Journal of Language and Communication Disorders*, *57*(3), 615–629. https://doi.org/10.1111/1460-6984.12712

MacWhinney, B., & Leinbach, J. (1991). Implementations are not conceptualizations: Revising the verb learning model. *Cognition*, *40*, 121–157. https://doi.org/10.1016/0010-0277(91)90048-9

Malisz, Z., Brandt, E., Möbius, B., Oh, Y. M., & Andreeva, B. (2018). Dimensions of segmental variability: Interaction of prosody and surprisal in six languages. *Frontiers in Communication*, *3*(25). https://doi.org/10.3389/fcomm.2018.00025

Masaki, S., Akahane-Yamada, R., Tiede, M. K., Shimada, Y., & Fujimoto, I. (1996). An MRI-based analysis of the English /r/ and /l/ articulations. *International Conference on Spoken Language Processing, ICSLP, Proceedings*, *3*, 1581–1584. https://doi.org/10.1109/icslp.1996.607922

Ménard, L., Toupin, C., Baum, S. R., Drouin, S., Aubin, J., & Tiede, M. (2013). Acoustic and articulatory analysis of French vowels produced by congenitally blind adults and sighted adults. *The Journal of the Acoustical Society of America*, *134*(4), 2975–2987. https://doi.org/10.1121/1.4818740

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *Proceedings of the International Conference on Learning Representations (ICLR 2013)*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (pp. 3111–3119). Curran Associates, Inc.

Moisik, S. R., Esling, J. H., Crevier-Buchman, L., & Halimi, P. (2019). Putting the larynx in the vowel space: Studying larynx state across vowel quality using MRI. *The 19th International Congress of Phonetic Sciences*, 1–5.

Moscoso Del Prado Martín, F., Kostić, A., & Baayen, R. H. (2004). Putting the bits together: An information theoretical perspective on morphological processing. *Cognition*, *94*(1), 1–18. https://doi.org/10.1016/j.cognition.2003.10.015

Müller, A. (2015). *Analyse von Wort-Vektoren deutscher Textkorpora* (Bachelor's Thesis). Technische Universität Berlin. https://devmount.github.io/GermanWordEmbeddings

Munson, B., & Solomon, N. P. (2004). The effect of phonological neighborhood density on vowel articulation. *Journal of Speech, Language, and Hearing Research*, *47*(5), 1048–1058.

Nayak, K. S., Lim, Y., Campbell-Washburn, A. E., & Steeden, J. (2022). Real-time magnetic resonance imaging. *Journal of Magnetic Resonance Imaging*, *55*(1), 81–99. https://doi.org/10.1002/jmri.27411

Nelson, W. L. (1983). Physical principles for economies of skilled movements. *Biological Cybernetics*, *46*(2), 135–147. https : / / doi . org / 10 . 1007 / BF00339982

Nieder, J. (2023). A discriminative lexicon approach to word comprehension, production and processing: Maltese plurals. *Language*.

Nittrouer, S., Studdert-Kennedy, M., & McGowan, R. S. (1989). The emergence of phonetic segments: Evidence from the spectral structure of fricative-vowel syllables spoken by children and adults. *Journal of Speech and Hearing Research*, *32*(1), 120–132.

Noiray, A., Ménard, L., & Iskarous, K. (2013). The development of motor synergies in children: Ultrasound and acoustic measurements. *The Journal of the Acoustical Society of America*, *133*(1), 444–452. https://doi.org/10.1121/1.4763983

Noiray, A., Wieling, M., Abakarova, D., Rubertus, E., & Tiede, M. (2019). Back from the future: Non-linear anticipation in adults and children's speech. *Journal of Speech, Language and Hearing Research*, *62*(8S), 3033–3054.

Öhman, S. E. G. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *The Journal of the Acoustical Society of America*, *39*, 151–168. https://doi.org/10.1121/1.1909864

Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. *The Quarterly Journal of Experimental Psychology*, *17*(4), 273–281.

Oppenheim, G. M., Dell, G. S., & Schwartz, M. F. (2010). The dark side of incremental learning: A model of cumulative semantic interference during lexical access in speech production. *Cognition*, *114*(2), 227–252.

Palo, P. (2019). *Measuring pre-speech articulation* (PhD thesis). Queen Margaret University.

Palo, P. (2020). Can we detect initiation of tongue internal changes before overt movement onset in ultrasound? *Proceedings of ISSP 2020 - 12th International Seminar on Speech Production*.

Palo, P. (2021). Computer assisted segmentation of tongue ultrasound and lip videos. *Acoustic Week in Canada*.

Palo, P., & Lulich, S. M. (2021). An ultrasound study of the effect of rest position on timing of pre-acoustic speech movements. *Proceedings of Meetings on Acoustics*, *45*.

Palo, P., Schaeffler, S., & Scobbie, J. M. (2014). Pre-speech tongue movements recorded with ultrasound. *Proceedings of the 10th International Seminar of Speech Production*, 304–307.

Perkell, J. S., Cohen, M. H., Svirsky, M. A., Matthies, M. L., Garabieta, I., & Jackson, M. T. T. (1992). Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *The Journal of the Acoustical Society of America*, *92*(6), 3078–3096. https://doi.org/10.1121/1.404204

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1-2), 73–193. https://doi.org/10.1016/0010-0277(88)90032-7

Plag, I. (1999). *Morphological productivity: Structural constraints in English derivation*. Mouton de Gruyter.

Plag, I. (2013). *The oxford reference guide to English morphology*. Oxford University Press.

Plag, I., & Ben Hedia, S. (2018). The phonetics of newly derived words: Testing the effect of morphological segmentability on affix duration. In S. Arndt-Lappe, A. Braun, C. Moulin, & E. Winter-Froemel (Eds.), *Expanding the lexicon: Linguistic innovation, morphological productivity, and ludicity* (pp. 93–116). De Gruyter. https://doi.org/10.1515/9783110501933-095

Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2005a). Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica*, *62*, 146–159. https://doi.org/10.1159/000090095

Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2005b). Lexical frequency and acoustic reduction in spoken Dutch. *The Journal of the Acoustical Society of America*, *118*(4), 2561–2569. https://doi.org/10.1121/1.2011150

R Core Team. (2022). R: A language and environment for statistical computing. https://www.r-project.org/

Ramscar, M. (2002). The role of meaning in inflection: Why the past tense doesn't require a rule. *Cognitive Psychology*, *45*, 45–94.

Rapp, S. (1995). Automatic phonemic transcription and linguistic annotation from known text with Hidden Markov Models: An aligner for German. *Proceedings of ELSNET Goes East and IMACS Workshop "Integration of Language and Speech in Academia and Industry"*.

Reimers, K., Reimers, C. D., Wagner, S., Paetzke, I., & Pongratz, D. E. (1993). Skeletal muscle sonography: A correlative study of echogenicity and morphology. *Journal of Ultrasound in Medicine*, *12*(2), 73–77. https://doi.org/10.7863/jum.1993.12.2.73

Repp, B. H., & Mann, V. A. (1982). Fricative–stop coarticulation: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America*, *71*(6), 1562–1567.

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American Psychologist*, *43*(3), 151–160.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). Appleton-Century-Crofts.

Roelofs, A. (1997). The WEAVER model of word-form encoding in speech production. *Cognition*, *64*(3), 249–284. https://doi.org/10.1016/S0010-0277(97)00027-9

Rossi, M., & Autesserre, D. (1981). Movements of the hyoid and the larynx and the intrinsic frequency of vowels. *Journal of Phonetics*, *9*(2), 233–249. https://doi.org/10.1016/s0095-4470(19)30938-6

Rubenstein, H., Garfield, L., & Millikan, J. A. (1970). Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior*, *9*(5), 487–494. https://doi.org/10.1016/S0022-5371(70)80091-3

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, vol 2: Psychological and biological models* (pp. 216–271). The MIT press.

Saito, M. (2020). Pyult: Preprocessing utilities for ultrasound data in Python. https://doi.org/https://doi.org/10.5281/zenodo.4022838

Saito, M., Tomaschek, F., & Baayen, R. H. (2021). Relative functional load determines co-articulatory movements of the tongue tip. *Proceedings of the 12th International Seminar on Speech Production (ISSP 2020)*, 210–213.

Sanders, I., & Mu, L. (2013). A three-dimensional atlas of human tongue muscles. *Anatomical Record*, *296*(7), 1102–1114. https://doi.org/10.1002/ar.22711

Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(1), 1–17. https://doi.org/10.1037/0096-1523.3.1.1

Schmidtke, D., Van Dyke, J. A., & Kuperman, V. (2021). CompLex: An eye-movement database of compound word reading in English. *Behavior Research Methods*, *53*(1), 59–77. https://doi.org/10.3758/s13428-020-01397-1

Schreuder, R., & Baayen, R. H. (1997). How complex simplex words can be. *Journal of Memory and Language*, *37*(1), 118–139. https://doi.org/10.1006/jmla.1997.2510

Sereno, J. A., Baum, S. R., Marean, G. C., & Lieberman, P. (1987). Acoustic analyses and perceptual data on anticipatory labial coarticulation in adults and children. *The Journal of the Acoustical Society of America*, *81*(2), 512–519.

Seyfarth, S., Garellek, M., Gillingham, G., Ackerman, F., & Malouf, R. (2017). Acoustic differences in morphologically-distinct homophones. *Language, Cognition and Neuroscience*, *33*(1), 32–49. https://doi.org/10.1080/23273798.2017.1359634

Shafaei-Bajestan, E., & Baayen, R. H. (2018). Wide learning for auditory comprehension. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 966–970. https://doi.org/10.21437/Interspeech.2018-2420

Shafaei-Bajestan, E., Moradipour-Tari, M., Uhrig, P., & Baayen, R. H. (2021). LDL-AURIS: A computational model, grounded in error-driven learning, for the comprehension of single spoken words. *Language, Cognition and Neuroscience*, 1–28. https://doi.org/10.1080/23273798.2021.1954207

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(4), 623–656.

Shaw, S. M., & Martino, R. (2013). The normal swallow: Muscular and neurophysiological control. *Otolaryngologic Clinics of North America*, *46*(6), 937–956. https://doi.org/10.1016/j.otc.2013.09.006

Slud, E., Stone, M., Smith, P. J., & Goldstein Jr., M. (2002). Principal components representation of the two-dimensional coronal tongue surface. *Phonetica*, *59*(2-3), 108–133. https://doi.org/10.1159/000066066

Smith, R., Baker, R., & Hawkins, S. (2012). Phonetic detail that distinguishes prefixed from pseudo-prefixed words. *Journal of Phonetics*, *40*(5), 689–705. https://doi.org/10.1016/j.wocn.2012.04.002

Song, J. Y., Demuth, K., Shattuck-Hufnagel, S., & Ménard, L. (2013). The effects of coarticulation and morphological complexity on the production of En-

glish coda clusters: Acoustic and articulatory evidence from 2-year-olds and adults using ultrasound. *Journal of Phonetics*, *41*(3-4), 281–295.

Steele, C. M., & van Lieshout, P. H. H. M. (2004). Use of electromagnetic midsagittal articulography in the study of swallowing. *Journal of Speech, Language, and Hearing Research*, *47*(2), 342–352.

Stein, S. D., & Plag, I. (2021). Morpho-phonetic effects in speech production: Modeling the acoustic duration of English derived words with linear discriminative learning. *Frontiers in Psychology*, *12*(678712). https://doi.org/10.3389/fpsyg.2021.678712

Stein, S. D., & Plag, I. (2022). How relative frequency and prosodic structure affect the acoustic duration of English derivatives. *Laboratory Phonology*.

Stolar, S., & Gick, B. (2013). An index for quantifying tongue curvature. *Canadian Acoustics - Acoustique Canadienne*, *41*(1), 11–15.

Stone, M. (2005). A guide to analysing tongue motion from ultrasound images. *Clinical Linguistics and Phonetics*, *19*(6-7), 455–501. https://doi.org/10.1080/02699200500113558

Stone, M., Davis, E. P., Douglas, A. S., NessAiver, M., Gullipalli, R., Levine, W. S., & Lundberg, A. J. (2001). Modeling tongue surface contours from cine-MRI images. *Journal of Speech, Language and Hearing Research*, *44*(5), 1026–1040.

Stone, M., Faber, A., Raphael, L. J., & Shawker, T. H. (1992). Cross-sectional tongue shape and linguopalatal contact patterns in [s], [ʃ], [f], and [l]. *Journal of Phonetics*, *20*(2), 253–270. https://doi.org/10.1016/s0095-4470(19)30626-6

Stone, M., Goldstein, M. H., & Zhang, Y. (1997). Principal component analysis of cross sections of tongue shapes in vowel production. *Speech Communication*, *22*(2-3), 173–184.

Strycharczuk, P., & Scobbie, J. M. (2016). Gradual or abrupt? The phonetic path to morphologisation. *Journal of Phonetics*, *59*, 76–91. https://doi.org/10.1016/j.wocn.2016.09.003

Strycharczuk, P., & Scobbie, J. M. (2017). Fronting of southern British English high-back vowels in articulation and acoustics. *The Journal of the Acoustical Society of America*, *142*(1), 322–331. https://doi.org/10.1121/1.4991010

Sugahara, M., & Turk, A. (2009). Durational correlates of English sublexical constituent structure. *Phonology*, *26*(3), 477–524.

Sung, J.-H. (2014). The articulation of lexical and post-lexical palatalization in Korean. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 1678–1682.

Taft, M. (1979). Recognition of affixed words and the word frequency effect. *Memory & Cognition*, *7*(4), 263–272.

Taft, M., & Forster, K. I. (1975). Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior*, *14*(6), 638–647. https://doi.org/10.1016/0022-5371(76)90054-2

Tiede, M., Mooshammer, C., Goldstein, L., Shattuck-Hufnagel, S., & Perkell, J. S. (2011). Motor learning of articulator trajectories in the production of novel utterances. *Proceedings of the XVIIth International Congress of Phonetic Sciences*, 1986–1989.

Tomaschek, F., Arnold, D., Bröker, F., & Baayen, R. H. (2018). Lexical frequency co-determines the speed-curvature relation in articulation. *Journal of Phonetics*, *68*, 103–116.

Tomaschek, F., Plag, I., Ernestus, M., & Baayen, R. H. (2019). Phonetic effects of morphology and context: Modeling the duration of word-final S in English with naïve discriminative learning. *Journal of Linguistics*, 1–39. https://doi.org/10.1017/S0022226719000203

Tomaschek, F., Tucker, B. V., & Baayen, R. H. (2019). How is anticipatory coarticulation of suffixes affected by lexical proficiency? *PsyArXiv (Preprent)*, 1–34.

Tomaschek, F., Tucker, B. V., Fasiolo, M., & Baayen, R. H. (2018). Practice makes perfect: the consequences of lexical proficiency for articulation. *Linguistics Vanguard*, *4*(s2), 1–13.

Tomaschek, F., Tucker, B. V., Ramscar, M., & Baayen, R. H. (2021). Paradigmatic enhancement of stem vowels in regular English inflected verb forms. *Morphology*, *31*(2), 171–199. https://doi.org/10.1007/s11525-021-09374-w

Tomaschek, F., Wieling, M., Arnold, D., & Baayen, H. (2013). Word frequency, vowel length and vowel quality in speech production: An EMA study of the importance of experience. *Proceedings of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH 2013)*, 1302–1306.

Tucker, B. V., Sims, M., & Baayen, R. H. (2019). Opposing forces on acoustic duration. *PsyArXiv*, 1–38. https://doi.org/10.31234/osf.io/jc97w

Turton, D. (2015). Determining categoricity in English /l/-darkening: A principal component analysis of ultrasound spline data. *Proceedings of the 18th International Congress of Phonetic Sciences*.

Van Son, R. J., & Van Santen, J. P. (2005). Duration and spectral balance of intervocalic consonants: A case for efficient communication. *Speech Communication*, *47*(1-2), 100–123. https://doi.org/10.1016/j.specom.2005.06.005

van Rij, J., Hendriks, P., van Rijn, H., Baayen, R. H., & Wood, S. N. (2019). Analyzing the time course of pupillometric data. *Trends in Hearing*, *23*, 1–22.

Vitevitch, M. S. (1997). The neighborhood characteristics of malapropisms. *Language and Speech*, *40*(3), 211–228.

Vitevitch, M. S. (2002). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning, Mem-*

ory, and Cognition, *28*(4), 735–747. https://doi.org/10.1038/mp.2011.182. doi

Vitevitch, M. S., & Luce, P. A. (2016). Phonological neighborhood effects in spoken word perception and production. *Annual Review of Linguistics*, *2*. https://doi.org/10.1146/annurev-linguist-030514-124832

Vitevitch, M. S., & Stamer, M. K. (2006). The curious case of competition in Spanish speech production. *Language and Cognitive Processes*, *21*(6), 760–770. https://doi.org/10.1080/01690960500287196

Wang, J., Green, J. R., Samal, A., & Yunusova, Y. (2013). Articulatory distinctiveness of vowels and consonants: A data-driven approach. *Journal of Speech, Language, and Hearing Research*, *56*(5), 1539–1551. https://doi.org/10.1044/1092-4388(2013/12-0030)

Wang, S. H., Keenan, B. T., Wiemken, A., Zang, Y., Staley, B., Sarwer, D. B., Torigian, D. A., Williams, N., Pack, A. I., & Schwab, R. J. (2020). Effect of weight loss on upper airway anatomy and the apnea–hypopnea index the importance of tongue fat. *American Journal of Respiratory and Critical Care Medicine*, *201*(6), 718–727. https://doi.org/10.1164/rccm.201903-0692OC

Westbury, J. R. (1994). *X-ray microbeam speech production database user's handbook (Version 1.0)*. Universtify of Wisconsin-Madsion.

Whaley, C. P. (1978). Word-nonword classification time. *Journal of Verbal Learning and Verbal Behavior*, *17*(2), 143–154. https://doi.org/10.1016/S0022-5371(78)90110-X

Widrow, B., & Hoff, M. E. (1960). Adaptive switching circuits. *1960 WESCON Convention Record Part IV*, 96–104.

Wieling, M., Nerbonne, J., & Baayen, R. H. (2011). Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLoS ONE*, *6*(9). https://doi.org/10.1371/journal.pone.0023613

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd). CRC Press.

Wrench, A., & Balch-Tomes, J. (2022). Beyond the edge: Markerless pose estimation of speech articulators from ultrasound and camera images using DeepLabCut. *Sensors*, *22*(3), 1–27. https://doi.org/10.3390/s22031133

Wrench, A., & Beck, J. (2022). Physiological foundations. In R.-A. Knight & J. Setter (Eds.), *The Cambridge handbook of phonetics* (pp. 11–39). Cambridge University Press.

Wright, R. (2004). Factors of lexical competition in vowel articulation. In J. Local, R. Ogden, & R. Temple (Eds.), *Phonetic interpretation: Papers in laboratory phonology vi* (pp. 75–87). Cambridge University Press. https://doi.org/10.1017/cbo9780511486425.005

Wurm, L. H., Ernestus, M., Schreuder, R., & Baayen, R. H. (2006). Dynamics of the auditory comprehension of prefixed words. *The Mental Lexicon*, *1*(1), 125–146. https://doi.org/10.1075/ml.1.1.08wur

Yu, J. L., Wiemken, A., Schultz, S. M., Keenan, B. T., Sehgal, C. M., & Schwab, R. J. (2022). A comparison of ultrasound echo intensity to magnetic resonance imaging as a metric for tongue fat evaluation. *Sleep*, *45*(2).

Zharkova, N., Hewlett, N., & Hardcastle, W. J. (2011). Coarticulation as an indicator of speech motor control development in children: An ultrasound study. *Motor Control*, *15*(1), 118–140.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley Press.