Finding Structure in Silence: A distributed, discriminative approach to structure and representation in spoken communication

D i s s e r t a t i o n

zur Erlangung des akademischen Grades

Doktor der Philosophie

in der Philosophischen Fakultät

der Eberhard Karls Universität Tübingen

vorgelegt von

Maja Linke

aus

Banja Luka

2023

Gedruckt mit Genehmigung der Philosophischen Fakultät
der Eberhard Karls Universität Tübingen

Dekan: Prof. Dr. Jürgen Leonhardt

Hauptberichterstatter: Prof. Dr. Michael Franke
Mitberichterstatter: Assoc. Prof. Steven Piantadosi, PhD

Tag der mündlichen Prüfung: **09.05.2023**

Universitätsbibliothek Tübingen, TOBIAS-lib

Finding Structure in Silence: A distributed, discriminative approach to structure and representation in spoken communication

---

Maja Linke

May 15, 2023
Version:

Abstract

Talk is a fundamental human experience, and speech signals are the first structured source of information we encounter in our environment. Yet the fact that talking is such a common and seemingly effortless activity easily obscures the challenges involved in explaining how it works. At a closer look, however, speech may be one of the most complex behaviors humans exhibit. It involves fine-grained kinetic adaptations that unfold in real-time and result in predictable, intelligible signals. Speech evolves through sensitivity to sensory prediction errors from multiple sources. It involves both co-ordinating one's own behavior (the way speech signals are articulated in context) and modeling of mutual expectations in time. All of these mechanisms rest on learning and develop across the lifespan. Given that speakers learn from exposure to speech samples that vary with experience, how do they ever manage to maintain sufficiently similar models of expectations?

In this dissertation, I investigate the sources of information (signals) that allow speakers to co-ordinate their expectations and successfully communicate. I show that regular patterns of co-occurrence between speech forms at various levels of abstraction serve as context to structure and manage the uncertainties of communication while gradually increasing the rate at which fine-grained differences in articulated signals are perceived. I propose that this leads to a predictable ebb and flow of uncertainty that allows us to maintain mutually predictable time templates and, therefore, a distributed transmission process that is, to some degree, experience-independent.

This work aims to explain how structure and form emerge from human vocal signals and to answer a fundamental question about the nature of the stable temporal organization of signals produced in human communication. It applies statistical and computational tools to transcripts and recordings of speech, exploring how speaker experience shapes speech structure at various levels of description. The dissertation links topics in learning theory, measurement and information theory, linguistics, philosophy, cognitive science, and neuroscience. It provides insight into the dynamics of alignment and the role of coordination in efficient information transmission.

As for the mot juste, you are quite wrong. Style is a very simple matter: it is all rhythm. Once you get that, you cant use the wrong words. But on the other hand here am I sitting after half the morning, crammed with ideas, and visions, and so on, and cant dislodge them, for lack of the right rhythm. Now this is very profound, what rhythm is, and goes far deeper than words. A sight, an emotion, creates this wave in the mind, long before it makes words to fit it; and in writing (such is my present belief) one has to recapture this, and set this working (which has nothing apparently to do with words) and then, as it breaks and tumbles in the mind, it makes words to fit it. But no doubt I shall think differently next year.

— **Virginia Woolf writing to Vita Sackville-West**
16 March 1926

Contents

1 Introduction

> *I've had nice letters from people regretting that my talks are above them, and others equally nice regretting that they are below; so hadn't I better pursue the even tenor of my own way?*
>
> — **E.M. Forster**

I started working on a thesis in the fall of 2017. My research focused on the examination of articulations and speech errors involving phoneme and morpheme 'exchanges' (Dell, 1984). I was interested in whether the probability of error would increase depending on the extent to which phonemes and morphemes are discriminated from each other by the structure of speech sequences. The hypothesis was that learning would lead to competition between phonemes and morphemes and the contexts they occur in, such that over time contexts would compete for phonemes, and phonemes would compete for contexts. This thesis starts with my first discovery: that all of my initial research questions were ill posed.

1.1 When things go wrong in speech: the research question in context

In the initial data analyses, it became apparent that the majority of words in conversational English are affected by some form of a 'morpheme' or 'phoneme' distortion. Notably, almost 60% of all words in the Buckeye Corpus of Conversational Speech (Pitt et al., 2005) deviate from the dictionary model (the 'canonical' form), often to the extent that they are unintelligible in isolation. This effect is often referred to as *reduction* (Gregory et al., 1999; Bell et al., 2003; Bell et al., 2009; Gahl et al., 2012; Aylett and Turk, 2004, e.g.), yet a closer inspection of the phenomena described by this term reveal that the processes that give rise to them are not particularly well characterized by the word 'reduction'. The most frequent transcription of the word *probably*, for example, is *p r aa b l iy*, and not *p r aa b ah b l iy* as supposed by the dictionary. The word *probably* also occurs as *p r aa b, f r ay, p r aa w ah v w iy* and *p r ah iy*. Some words are almost always compressed, such that far from being

'reduced,' it would seem that their dictionary forms, in fact, represent an expansion. Meanwhile, although some words and utterances tend towards shorter (or reduced) forms, many of them do not. At both word and utterance levels, speakers often insert acoustic-phonetic contrasts, filled pauses, and conjunctions, such that the phenomenon appears to be best described as speech form variation. For example, the 300 000 token Buckeye corpus contains 313 transcribed forms of the word *that*. The words speakers produce when they mean to say *dh ae t*[1], are transcribed as

> *dh ih t, dh ae, dh eh dx, dh ih tq, n ae tq, n ae t, dh eh, dh ah tq, dh ah t, dh ih dx, ae tq, dh ih, th ae tq, n ae dx, eh tq, n ae, dh ae d, ae t, th ae t, dh ah dx, n eh t, dh eh d, d ae tq, ae, dh ah, eh t, eh, d ae t, n eh tq, dh ih d, th ih t, ae dx, z ae t, ih t, l ae tq, th eh t, ah, n ih t, ih tq, ih, th eh tq, ah tq, z eh t, dh ae k, z eh dx, z ae tq, n eh, ih dx, d eh t, ah t, z ih t, th eh dx, l ae t, dh ae p, z ah t, n eh dx, n ah t, n ae d, eh dx, dh ae b, d eh tq, ah dx, z ih, th ih, s eh tq, s ih t, m ae t, l uw, l ih tq, l ih dx, l ih, l eh s, l eh k, l ah dx, l ah d, l ae tq d, l ae d, k eh tq, k dh ae tq, k d ae t, k ae t, jh ah, iy t, ih t dh ae tq, ih ah dx, hh eh tq, hh eh d, hh ah, hh ae t, hh ae d, h ae t, f ae tq, f ae t, en n ae t, eh n, eh k, eh dh ae dx, dx ih tq, dx ih, dx ah tq, dx ah, dx, dh uw t, dh uh dx, dh uh d, dh t, dh l eh t, dh iy tq, dh ih zh, dh ih w, dh ih tq d, dh ih t dh, dh ih m, dh ih ah, dh er, dh eh z, dh eh dx ih, dh eh ae tq, dh dh ae dx, dh ay, dh ah n, dh ah dh, dh ah b, dh ae tq n, dh ae t d, dh ae s, dh ae m, dh ae eh tq, dh ae dx ah, dh aa tq, dh aa dx, d uh t, d eh b, d ah dx, d aa tq, ay tq, ay ay, ay, ah m, ah g, ae w, ae s, ae b tq, ae b, aa tq, aa dx, ...*

Many of them do not even resemble the dictionary model of the word *that*; the articulations are highly productive and seem to vary randomly. This raises one of the first questions we will address in this thesis: what *exactly* is it that speakers reduce or swap when they are described as 'exchanging'[2] phones and morphemes? **Do speakers really model their articulations on dictionary forms, so that we can assume that the forms they produce when they actually speak represent some kind of performance error, each being an involuntary deviation from the dictionary form they actually meant to produce?** One reason to doubt that this is the case are the well-documented struggles of text-to-speech synthesis to produce intelligible, natural-sounding speech sequences modeled on dictionary transcriptions (Duffy and Pisoni, 1992; Wester et al., 2016). Text-to-speech synthesis results in interpretable sequences, yet the cognitive strain involved in listening to text-to-speech voices of *Microsoft Mikes*[3] and *Microsoft Marys* makes their messages less

[1] which is the phonetic transcription of the 'canonical' form of the word *that*; the second most frequent form in the corpus, occurring 727 times, as opposed to its 'reduced' variant *dh ae tq*, which occurs 1085 times and is realized as *tha?*, with the final consonant *t* formed as an audible release of air after a full closure of the vocal folds in the larynx (a glottal stop)

[2] articulatory variations that lead to 'swapping' of phonemes and morphemes at the boundaries of consecutive words are frequently described as segment exchanges in the speech error literature

[3] one of the default text-to-speech voices delivered with the Microsoft operating system

memorable and harder to attend to (Paris et al., 2000). **Is speech form variation a bug or a feature?**

The literature suggests that many instances of what might appear to be 'articulatory noise' may actually serve as informative and productive parts of communicative sequences (e.g. Bell et al., 2003; Aylett and Turk, 2004; Bell et al., 2009; Wedel et al., 2013b; Seyfarth, 2014; Seyfarth et al., 2016; Wedel et al., 2018; Hall et al., 2018; Priva and Jaeger, 2018). That is, while deviant forms are often thought to be redundant in context, many of them seem to improve recall and intelligibility of speech sequences (Bosker et al., 2013; Arnold et al., 2003; Fraundorf and Watson, 2011) and be bounded by expectations (Schachter et al., 1991; Bosker et al., 2014a). Simultaneously, many of the forms produced by speakers are unattested even in dictionaries of 'form variance', and the extent to which individuals use them vary widely (cf. Byrd, 1994; Dilts, 2013; Bürki, 2018). That is, although many articulated form variants are never, by definition, 'canonical,' they nevertheless appear to be predictable enough to allow speech sequences to not only be understood but even to sound natural. This leads to a second question: Are *speech errors* communicative? Which raises a third: What is the difference between *forms* and *errors*?

These questions led me to reconsider my initial questions and approach. To address these new questions, we will first focus on the following: how might one formalize the distinction between errors (or noise) and structured, predictable variations (or signal). Simultaneously, this led me to consider a second, related question: if articulatory variants are signals, then how do we explain sources of information that structure them? To address these questions we will conduct a series of corpus analyses and computational simulations that revealed an interplay between the temporal structure of articulations in context (contextual uncertainty) and speech form production. The theoretical considerations and the rationale for modeling choices constitute the conceptual core of this work. The quantitative analyses are theoretically motivated.

In what follows, I will introduce the theoretical context in which we will address these questions, relating them to a range of cross-disciplinary findings that broadly combine to form a distributed, discriminative theory of learning and communication (Ramscar, 2019; Ramscar, 2021a).

1.2 To err is human; to study error-making is cognitive science

When is a morpheme or phoneme exchange an error? Minimal word pairs, such as, for example, *fan* and *van*, are distinguished by a difference in the word-initial acoustic-phonetic contrast (a phoneme exchange). Errors, on the other hand, are often discriminated from words (or, more generally, forms) by context. To claim that one has *parked their fan behind the house* is an error that involves a phoneme exchange; Unless a wheeled fan was deposited behind a house. Then it is not, then it is just unusual. While distinguishing forms from errors is relatively easy when exchanges produce nonsense (i.e. unattested forms) or violate phonotactic conventions (which they rarely do), some exchanges result in messages that are less common and, at the same time, interpretable (puns). Classifying these involves both knowledges about the context and the knowledge about the individual speaker (what she might have **not** meant to say). The problem is not new (inter alia Grice, 1969; Clark and Brennan, 1991; Prince, 1981); how do we approach it?

We shall start with the message. In their essay *To err is human* Hofstadter and Moser (1989) reminisce on years of speech error-collecting *as a hobby and a serious activity*. Their compendium introduces a rough classification scheme that covers inappropriate speech in its various forms: malaproprisms[4], spoonerisms[5], mixed and otherwise awkward metaphors. In the onset they observe (Hofstadter and Moser, 1989, p. 2):

> On first thought, it might seem surprising that it takes practice to become a good collector of errors, but when one begins to try to identify and classify specific kinds of mistakes, one sees how hard it is to pinpoint or remember them amidst the constant swirl of language. Indeed, most speech errors go completely unnoticed by. Both to speaker and hearer unless someone points them out. A typical listener hears just the *content* of the utterance without noticing that something has gone awry with its *form*. The reason for this is that most errors are not simply random intrusions of "noise" into an otherwise clear and unambiguous flow of communication; they are almost always intimately connected with the

[4] Malaproprisms (or dogberryisms) are involuntary word-exchanges where the substitutes and targets are easily mixed-up because they sound similar and form bizarre or funny utterances. They are termed after famous characters from theater plays Mrs. Malaprop (The Rivals) and Dogberry (Shakespeares' *Much Ado about Nothing*) who were prone to saying things like *illiterate him quite from your memory* (instead of *obliterate*).

[5] Spoonerisms involve the earlier mentioned involuntary exchanges of corresponding speech sounds or morphemes between two consecutive words in a phrase which leads to sentences such as *Is the bean dizzy?*, instead of *Is the dean busy?*.

> speaker's intended message, and reveal something of it. Rather than blatantly standing out from the rest of the utterance, a typical error blends in smoothly with it.

A similar phenomenon is evident to anyone who has ever attempted to transcribe spontaneous speech. Deviation from the 'target' form tends to not stick out and instead blends in smoothly with the context. Not only do more than half of the words in connected speech fail to resemble target forms defined by dictionaries, but simultaneously, the sounds, forms, and words that transcribers 'hear' in acoustic signals will vary considerably across individuals (Raymond et al., 2002). In fact, even trained individuals' estimates deviate when words are presented out of context, and the extent to which they do seems to vary with the amount of contextual information provided by the surrounding words (Ernestus et al., 2002). This indicates that 'form acceptability' is a function of context. If the form is defined by the context, what *exactly* is a context?

Context can be defined as the organizing principle that subserves efficient management of speaker expectations. One approach to modeling context developed from the study of word distributions in large text resources and builds on the observation that a word's meaning can be inferred from the lexical neighborhoods it tends to appear in (McDonald and Ramscar, 2001; Gleitman, 2002). Distributional approaches define context in terms of collocational regularities in word sequences. Shared collocational regularities tend to coincide with shared semantic dimensions (Harris, 1954; Lund and Burgess, 1996; Landauer and Dumais, 1997; Sahlgren, 2008).

The distributional context here should not be mistaken with its computational implementations, a probability of the word in a specific lexical frame it is observed in; in terms of **learnable speaker expectations**, a distributional context describes a cluster of words that distributional regularities discriminate from all other words, but not from each other. This description of context is motivated by learning theory. Ramscar (see 2019, p.41) introduces it as follows:

> Because learning including classical conditioning is best characterized as a predictive discriminative process (Ramscar et al., 2010; Ramscar and Port, 2016), it follows that if two or more words have the same conditioning histories (that is, if the vectors of their co-occurrence patterns in relation to other words are identical), then while someone exposed to this. distribution will learn to discriminate these words from the words in the rest of the lexicon that dont share the same conditioning history, she wont learn to discriminate them from one another. Further, because

> learning is a probabilistic process (i.e., the degree to which learners will come to discriminate the expected behavior of one word from that of another will be a matter of degree) where two or more words have conditioning histories that vary only slightly from one another, but greatly from other words, a learners expectations about the behavior of the words in such a set will be far less discriminated within set than they will be from the rest of the lexicon, such that members of the set will tend to cluster in the lexicon.

This definition is, in its essence, systemic. It seems to gives cues to the dynamics of distributional structures that can vary across speakers and vary in time. In other words, lexical regularities[6] appear to support speech recognition by regulating and structuring the information in and thus predictability of signal sequences (Blevins et al., 2016; Ramscar et al., 2018). This suggests, at least on the surface, that one possible source of input to the process of signal/noise discrimination could be lexical and grammatical forms themselves. However, this observation simply serves to expose a seemingly paradoxical aspect of speech, namely that the realization of words and their component parts are not only often ambiguous, but in fact, that forms speakers recall having heard are often not present in the speech signal and can only be inferred in context (Port and Leary, 2005; Ernestus et al., 2002). Moreover, this 'context' is not merely defined by the presence or absence of articulated signals (and their perceptual correlates) but also by the timing of these signals (Dilley and Pitt, 2010; Morrill et al., 2014; Baese-Berk et al., 2019). The acoustic phenomena related to signal perception and segmentation will be introduced in more detail in chapters 5 and 6.

Most recent approaches to speech assume it to be a probabilistic process that involves extracting forms from a noisy signal (Clayards et al., 2008; Kleinschmidt and Jaeger, 2015). It follows, accordingly, that if the identification of the 'contents' of the speech signal is a probabilistic process, then it must somehow rely on the alignment of speaker expectations. That is, statistical regularities in the speech signal can only be noticed if speakers have already managed to extract some shared structure(s) from the signals they have previously been exposed to (and learned from). This suggests that as far as the speech signal goes, the evidence that would allow language users to identify forms appears to be insufficient to provide a basis for the initial structuring or synchronization of speaker expectations. That is, the assumption that regularities at the form/lexical level might serve to somehow facilitate the alignment of speaker expectation ignores the fact that the identification of 'items' at the form/lexical level relies on the existence of shared expectations in order for these items to be extracted/inferred in the first place. This is, of course, is a chicken

[6] i.e., regular patterns of co-occurrence between words

and egg problem: you cannot detect forms in the signal without shared statistical structure, and you cannot extract/learn statistical structure if you cannot first detect the forms.

1.3 The 1000 (or more) minds problem in signal/noise discrimination

The problem of discriminating speech forms from speech errors is essentially the problem of signal/noise discrimination. This problem is best understood in terms of information theory (Shannon, 1948). In very general terms, information theory can be viewed as a theory that defines signals or information in terms of predictable variation. In contrast to the theory, where signals are defined by fixed codes shared by transistors, human communicative codes are learned from exposure. In addition, in speech, unlike in information theory, code words are time-varying quantities: the boundary of a 'code word' is set by the rate at which speech signals (puffs of air modified by articulator configurations) unfold. Words, syllables, and speech sounds represent predictable fluctuations in frequency, intensity, and the rate at which more or less predictable (informative) changes in the signal occur. Because speakers' expectations (about which part of the acoustic signal is predictable and which is not) vary with the context and speaker experience, which part of the transmitted quantities can serve signals will also vary.

Changes in signal dimensions that are not predictable cannot serve as communicative signals. Stretches of English or Bulgarian, for example, can sound like noise or babble enhanced by prosody to a person who has not learned English or Bulgarian. Language learning allows language users to gradually identify predictable, functional patterns of vocal variation that give form to speech signals: changes in duration and the acoustic properties, phrases, words, syllables, and sounds. From this perspective, form is defined by learning to share communicative expectations. The human mind can theoretically impose structure on any given source of regularity. This seems like a problem, given that the acoustic signals speakers are exposed to vary in almost every aspect. If most words, syllables, and phonemes do not share many of their acoustic features and those that they share occur at varying rates, how can 1000 (or more) speakers learn them from exposure? How do speakers manage to acquire and maintain similar expectations about which parts of what they hear constitute noise and which parts of it constitutes a signal? How is all of this learned?

The discriminative theory of communication (Ramscar and Port, 2016; Ramscar, 2019; Ramscar, 2021a) uses the error-driven learning paradigm to explain how

children and adults acquire and optimize their language models through learning from the latent structure of signals they are exposed to. It highlights the role of developmental constraints and cumulative experience in shaping learning and behavior over time. Discriminative learning was used to explain the order in which children extract regularities from sequences, over-generalization in early language learning, why children are better at learning aspects of languages that adults find very hard to learn (e.g. gender and irregular plurals), and how early learning and brain development change speakers experience of information (aspects of linguistic signals they attend to) (Ramscar and Yarlett, 2007; Ramscar et al., 2013b; Ramscar et al., 2013c). Moreover, interactions between the linguistic structure and domain general learning mechanisms explain how experience affects peoples' linguistic performance across adulthood (Ramscar et al., 2014), and in multilingual adults (Ramscar et al., 2017). Critically, the discriminative approach introduces a dynamic notion of 'structure' and 'representation,' where the signal resolution develops with experience. The informativeness of morphological regularities, words, or multi-word phrases changes with learning, becoming increasingly fine-grained (i.e. better discriminated) where necessary and increasingly smudged where the communicative task does not require for it.

The discriminative learning model can be summarized as follows: learning shapes behavior, and behavior reflects the structure of the experienced environment. In communication, the behavior altered by learning contributes to the structure of the environment (speakers speak). Because learning is discriminative – learners acquire new knowledge in relation to old knowledge by noticing informative features that discriminate new from old knowledge – learners' models of the world and their knowledge representations will become increasingly diversified and detailed over time. This necessarily leads to more divergent patterns of behavior and to more variability across learners at different levels of experience. At the same time, those aspects of the communicative environment that are relatively stable in time will become increasingly uninformative. Speakers will learn to not attend (consciously) to these 'overpredicted' aspects of the environment. This will decrease speakers' sensitivity to the variability in overpredicted forms. The systematic convergence of language users' expectations about certain aspects of the linguistic codes and developmental constraints (e.g., the protracted development of the prefrontal cortex, more in section 6) will impose limits on learning and the structure of the models that arise as a consequence of this learning. This suggests that the maintenance of learnable structures is imposed by constraints to learning (what can be learned). The details of the theoretical model and the concrete predictions derived from it in this work will be introduced further in chapter 2.

1.4 Synopsis

Given the preceding, the empirical and theoretical focus of the thesis is on the statistical structure of spontaneous speech. It examines the hypothesis that the statistical structure of spoken signals (i.e., the structure of signal distributions) reflects functional requirements of probabilistic communication. Two competing requirements determine the function: on the one hand, languages are learned and must be sufficiently structured to support learning from input, and on the other, languages are used to communicate and must possess a structure that allows speakers' to maintain sufficiently similar expectations independent of their experience. This work aims to test the functional hypothesis by examining distributions of phrases, grammatical forms, words, and word variants in conversational English.

The thesis is laid out in 8 chapters that span the literature synthesis and the analyses that helped me redefine the research question, chapters 3 to 6, which form the core of this work, and closes off with chapter 7, where we extend findings on the 'structured distribution of uncertainty' to a model of visual word recognition. In what follows, I shall introduce the content of the individual chapters and outline the questions they serve to address.

Chapter 2 synthesizes the exploratory analyses and literature that informed the investigations reported in this thesis. The chapter starts off introducing the linguistic perspective on the process of articulation and form variation in conversational speech. It establishes the linguistic notion of form in the context of its (neuro)physiological underpinnings, its social function and the special status it has among the variety of communicative behaviors.

Throughout the chapter, we highlight the effects the distributed nature of human habits and conventions has on the temporal structure of human behaviors (including human communicative behaviors), and the quantifiable changes in the statistical structure of aggregated data samples this involves. The section introduces the methods and the terminology used throughout the thesis. It establishes the theoretical definition of information and the challenges modeling speech signals as information involves. In particular, this section provides some background on the notion of information as a result of learning. It unveils problems that modeling a signal shaped by learning involves, highlighting the fact that learning outcomes vary with the structure of the input/environment, and that this input/environment (i.e. speech) is shaped by complex, multiplicative processes that change across multiple timescales. **Chapter 2**, *Finding answerable question*, serves to organize a range of empirical phenomena in the context of distributed learning, revealing some of the complexities involved in modeling speech.

Speech is a probabilistic process. Since the idea of probabilistic communication suggests that speakers have to be able to align their expectations about probabilities, understanding how they do this seems crucial. **Chapter 3**, *Speech form distributions in conversational English*, addresses the question of how speakers organize speech sequences to maintain mutual predictability in spoken communication. The chapter presents results from sampling simulations, alongside distributional analyses performed on the Buckeye corpus of conversational English. We discuss the findings in the context of learning and functional requirements of communication. The analyses seek to illuminate the organizing principles of communicative sequences that lead to formation of abstract communicative conventions on the example of grammar. We show how patterns of co-occurrence discriminate between functional roles that classes of words carry out in communication. Our results suggest that seemingly random variation in conversational speech optimizes speech contrast distributions for efficient, sample-invariant transmission at all levels of description. Taking verbs, nouns and function words I discuss how these functional roles can interact with word order and the average sequence position to maintain mutual predictability. The effects this has on the lexical productivity and subcategorization patterns of different lexical categories are discussed. The findings presented in this chapter show that context, which is set by systematic variation between linguistic cues at multiple levels of description, itself follows a distribution. These results suggest that power laws observed in word frequency distributions are a product of aggregating over a distribution of structurally distinct distributions of communicative cues that evolve on different timescales through learning. The content of this chapter is based on material published in Linke, M., and Ramscar, M. (2020). *How the probabilistic structure of grammatical context shapes speech*. Entropy, 22(1), 90.

Chapter 4, *From Information Structure to Speech Form*, extends the findings presented in chapter 3. It presents additional analyses of articulated word and phonetic segment variants from the Buckeye Corpus of conversational speech. The findings presented in chapter 3 raise questions about the source of structure in the distributions of seemingly random acoustic-phonetic variants. Despite the fact that parts of speech differ dramatically in both numbers of lexical types they host and rates at which these types reoccur, word initial speech segments across these categories converge in almost identical distributions. What gives shape to such structured randomness? Current information-theoretical accounts explain speech form variation as a function of word predictability. More frequent words tend to be more 'reduced' in articulations. Chapter 4 provides evidence that the number of articulated variants in the corpus is far better explained by a words contextual dispersion (the number of lexical collocations a word is observed in) than word frequency. This suggest that speech form variation might not be a function of redundancy, as suggested by most recent information-theoretical accounts. Instead, these results indicate that articulated forms vary systematically with the uncertainty of the context. Are

deviations from the dictionary form context-specific forms in their own right? Are they learned? Parts of the results in this chapter were presented in materials published in Linke and Ramscar, 2020a.

Chapter 5, *Finding structure in silence*, follows up on the question raised by the preceding two chapters: Which part of the speech signal can provide a shared source of information to constrain and structure speakers' expectations about acoustic dimensions that vary across speakers and contexts? The chapter introduces a theoretical rationale for the hypothesis that speech pauses play a crucial role in systematic temporal structuring of speech signals that result in systematic form distributions observed in chapter 3. It approaches the problem of alignment as a task of learning models that maintain shared expectations across multiple timescales. These models are learned from exposure to a noisy signal and structured by generations of speakers. Chapter 5 presents arguments why a memoryless source of information is a necessary precondition to counter the misalignment of expectations that this kind of complex distributed learning necessarily entails. The results presented here support the idea that pauses provide a memoryless source of information that facilitates alignment between speakers at different levels of experience through predictable interactions with articulation rates. Why focus on pauses and articulation rates?

The intelligibility of speech signals relies on entrainment – the ability of speakers to synchronize the rates at which informative changes in speech are transmitted and processed. The work presented in this thesis proposes that entrainment and the alignment processes that follow from entrainment are achieved through the statistical structure of spoken signals, and in this chapter shows how pauses offer a time-invariant template for structuring speech sequences that is available to speakers at all levels of experience. An analysis of corpus data taken from conversational Korean and English shows that pauses in both languages approximate the exponential distribution. The chapter describes how this memoryless distribution of pause aggregates can facilitate both the initial structuring and maintenance of predictability in spoken signals over time, and shows how the properties of this signal change predictably with speaker experience. The findings indicate that speaker experience leads to predictable changes in the way pauses interact with the morphological and rhythmical structure of languages, allowing speakers at all stages of lifespan development to distinguish signal from noise and maintain mutual predictability in time.

Material presented in this chapter were published on arXiv under arXiv:2112.08126 as Linke, M., and Ramscar, M. (2021). Finding Structure in Silence: The Role of Pauses in Aligning Speaker Expectations. and is currently under review at the journal of *Language, Cognition and Neuroscience*.

Chapter 6, *Morphology? How the distribution of uncertainty shapes speech form* completes the core part of the thesis, which deals with sources of systematic variation in speech signals. This chapter examines the effects of lifelong learning on the way forms are signaled in context. It puts forward arguments and analyses to support the proposal that lifelong learning leads to predictable changes in the the way older speakers form words in sequences, with a particular focus on the way verbs and nouns are articulated in relation to the duration of the preceding pause. The findings presented in chapter 5 and sampling simulations suggest that lifelong experience leads to systematic changes in the distribution of information (and uncertainty) across speech sequences and lexical clusters. These findings suggest that lifespan learning will lead to predictable differences in the extent to which speaker experience increases the uncertainty preceding English verbs and nouns. The results of analyses presented in this chapter indicate that speakers manage the shift in the distribution of uncertainty by adapting the signals they produce at the word boundaries, showing that the likelihood of segment deviation changes systematically as a function of uncertainty in the more lexically productive class of nouns, but not in verbs. The chapter relates these results to the differences in the morphological structure of English nouns and verbs, discussing how morphological form can arise from structure as a consequence of functional pressures.

The final **chapter 7**, *It's all in the code*, extends the principle of information rate and the distribution of uncertainty in time to spatial distributions of visual cues in words and random letter sequences. An analysis of the spatial distributions of low-level visual features shows how visual cue distributions in orthographic codes are at different levels of description structured to allow for both discriminating between words and non-words and discriminating words from other words. The implications of these results are discussed in relation to the research question and conclusions drawn from psycho-linguistic experimentation. The results of the analyses suggest that the representational status (form) in any communicative behavior, including speech, can only be interpreted in the context of the functional requirements of the communicative task and the individual who performs it.

Chapter 8 summarizes the conclusions of the work presented in this thesis and their significance for future work. The implications of these findings on the models, methods and metrics used in analyses of language data are discussed.

2

Down the Rabbit Hole: How distributed, discriminative speakers behave in time

> *People believe a little too easily that the function of the sun is to help the cabbages along.*
>
> — **Gustave Flaubert**

This chapter addresses the implications of distributed, discriminative learning on the organization of speech models and speech sequences in time. It exposes the similarities and differences between the approach taken in this work and other systemic and information-theoretic research on speech form variation. In what follows, we will summarize the predictions derived from previous work on the discriminative theory of communication (Ramscar and Port, 2016; Ramscar, 2019) and make explicit the contributions developed in the present work. In section 2.1.1, we will summarize the working assumptions at the core of the analyses presented in this work and the hypotheses that follow from them. Finally, section 2.1.2 relates the way in which the distribution of uncertainty across grammatical phrases, words, and morphemes changes with learning across adulthood with changes in the information rates and the temporal resolution of sequences. We discuss the implications of lifespan development on the relative frequencies of words from different frequency registers and the informativeness of words and segments from different registers to articulated signals.

This chapter emphasizes the importance of temporal constraints and learned expectations in the organization of human communicative sequences. It presents a summary of initial findings on long-standing questions about learning and the functional forces that shape word frequency distributions, speech signal resolution, and their relationship to the morpho-syntactic structure of languages.

2.1 Information in human communicative systems

The approach we take here views human communication as probabilistic and languages as systems of communicative conventions continuously shaped by the functional requirements of usage. In this, it is related to a large body of recent research that seeks to describe and explain linguistic regularities in terms of evolving information systems that are shaped by the functional requirements of efficient transmission (Cancho and Solé, 2003; Hale, 2003; Hale, 2006; Levy, 2008; Jaeger and Tily, 2011; Piantadosi et al., 2011; Piantadosi et al., 2012; Fedzechkina et al., 2012; Futrell et al., 2015; Cancho, 2017; Gildea and Jaeger, 2015; Dautriche et al., 2017; Cancho, 2017; Mahowald et al., 2018; Gibson et al., 2019; King and Wedel, 2020).

Among the individual strands of information-theoretic approaches to languages, there seems to be a relatively high degree of consensus on what the functional requirements of efficient communication are: articulated signals ought to help listeners minimize the uncertainty about messages gradually (or smoothly) by optimizing the sequential distribution of information (e.g., Bell et al., 2003; Aylett and Turk, 2004; Levy, 2008; Mahowald et al., 2013), while allowing speakers to produce signals with minimal effort (Zipf, 1949). Human communication involves efficient signaling [1]. How do humans achieve this?

While efficiency appears to be a shared core assumption underlying information-theoretic accounts of human communication, the details of its implementation are often focused on linguistic formalisms and formal, statistical process descriptions. While the former seems to assume the unambiguous existence of linguistic forms (or discrete events), the latter tend to rely on a particular distribution of discrete events (Shannon, 1948), or a particular distribution of distributions of discrete events (e.g., Blei et al., 2003)[2]. Meanwhile, the assumptions about the mechanisms involved in

[1] It is perhaps worth mentioning here that there are findings which suggest that principles of efficient communication apply to signals produced by other animals too (Cancho et al., 2013). Interestingly, Clink et al. (2020), who examined the communicative efficiency of male gibbon solos, report mixed results indicating that communicative efficiency in male gibbons may depend on the unit of analysis. Relatedly, investigation of the seasonal fluctuations in the structure of humpback whale songs raises questions on the status of information theoretical quantities in animal communication: humpback whales, who are well-known for notoriously inefficient singing bouts that can last up to 20 hours, appear to exhibit seasonal fluctuations in information-theoretical quantities. Furthermore, mixed analyses of units produced by humpback whales reveal gradual morphing of units along spectral, temporal, and spectro-temporal dimensions, maintaining the continuity of spectral content across subjectively dissimilar unit types (Mercado III and Perazio, 2021; Mercado and Perazio, 2022). Mercado and Perazio (2022) suggest *Given that it is not yet possible to experimentally identify the categories of units that are most salient from a humpback whales perspective, it is important to closely consider the acoustic variations that singing humpback whales produce, and to take those characteristics fully into account when analyzing songs.*

[2] Conceptually, information theory assumes a homogeneity among distributions. Bayesian methods, such as for example topic modeling, suppose heterogeneity among distributions, but implicitly

the emergence and evolution of linguistic representations (the perceptual correlates of discrete events) tend to be left underspecified. What are the mechanics of the efficient response in human communication? What is driving efficiency?

Recent empirical evidence suggests that speech forms are shaped dynamically by global communicative pressures. Examinations of gradual changes in the speech signal resolution suggest that these may be driven by the functional load on discriminative contrasts provided by minimal pairs – pairs of words discriminated by small differences in articulation (e.g. *bat* and *pat*). For example, Wedel et al. (2013b) present evidence that the probability of systematic alternation in vowel articulations that can lead to vowel mergers (where initially distinctly pronounced vowels eventually become indistinguishable) is negatively correlated with the number of minimal pairs that the merger would create in a context. Systemic approaches to speech signal production provide evidence that discriminative features are more likely to be maintained and 'hyperarticulated' if they provide information that allows speakers to distinguish between competing forms in context. Meanwhile, acoustic contrasts that are more likely to be distinguished from each other by the context they occur in are often 'hypoarticulated' and gradually vanish, leading to homophony[3] (Wedel, 2012; Wedel et al., 2013b; Wedel et al., 2018; Wedel et al., 2019b). This principle appears to hold across languages and seems to be shaped by incremental processing: across languages word-initial phonetic contrasts tend to be more diverse (Wedel et al., 2019a) and phonemic neutralization (variation that can lead to homophony, or loss of discriminative contrast) tends to be more common in word-medial and word-final positions [4].

Functional loads are mediated by context and seem to affect word boundaries differently. How exactly are functional pressures expressed in speakers' environments, and how exactly do the talking populations manage to respond to them with 'efficiency'? Winter and Wedel (2016) suggest an exemplar-based model in which phonetic form variation occurs in response to functional pressures triggered by increasing numbers of words[5]. The exemplar-based models suppose that ar-

(implicit in the way design choices are made) seem to assume homogeneity among models of distributions). These conceptual choices in models seem to be a relevant source of constraint with respect to the research questions language sciences seek to examine.

[3] a process by which words lose discriminative contrasts, such that hardly distinguishable words are used to mean different things

[4] We note here that the validity of these findings is somewhat controversial (see Sampson, 2013; Sampson, 2019, for critical discussion)

[5] Exemplar-based variation developed in response to abstraction-based models (Dell, 1984; Levelt, 1993), which stated in simplest terms assume a production process that relies on abstract units that can be stored at multiple levels (e.g., phonological or phonetic). The abstraction-based accounts assume that words have *underlying forms* that are stored in memory and that *surface forms* (i.e., the words people produce in natural conversations) are produced when speakers add, substitute or remove phonemes in a process guided by fixed rules.

ticulated variants are stored as 'memory-traces' or 'acoustic traces'[6]. Support for memory-based representations of variable forms mostly comes from psycholinguistic experiments that implement the shadowing paradigm, in which subjects repeat speech sequences after a short delay to the onset of hearing. Findings from these experiments suggest that speakers tend to imitate the variation in the phrase they hear (see, e.g., Goldinger, 1998). Further, patterns of variation in regional dialects seem to support the exemplar-based accounts, revealing consistent, convergent patterns of sub-lexical variation across speakers (Babel, 2010; Clopper and Pierrehumbert, 2008). This suggests that at least some of the variability in speech forms is learned.

There are, however, reasons to doubt that representation-based accounts of variation in speech can completely explain the empirical phenomena. First, imitation behaviors do not necessarily imply the existence of exemplars or representations (cf. Linke et al., 2017). In fact, signal discrimination does not seem to require that events be represented. How well people learn to discriminate between auditory stimuli appears to depend more on the variability of the input rather than the actual differences between the stimuli (Amitay et al., 2005; Amitay et al., 2006). That is, acoustic stimuli that do not differ in terms of the realized acoustic features are perceived as different in variable contexts (under uncertainty)[7]. Speakers do not need to memorize instances of variable representations to be able to align their articulations (i.e., produce signals). Instead, short-term dynamical alignment occurs on nearly all levels of description and appears to be related to the neural dynamics of temporal synchronization in natural speech. Speech segmentation in spontaneous speech relies on a process called entrainment, in which the activity in the auditory cortex 'shadows' the temporal and acoustic structure of speech signals through adaptation of neural oscillations[8]. Accordingly, at least some of the variability in the acoustic realization of speech signals appears to arise as a consequence of spontaneous adaptations to the temporal structure of speech signals. Moreover, acoustic adaptation often involves adaptations to the distribution of all acoustic contrasts rather than the adaptation of specific instances of the signal (Xie and Myers, 2017; Xie et al., 2017; Arbesman et al., 2010). The structure of acoustic signals appears to co-vary with time.

[6] What memory traces are, exactly, is, again, not entirely clear, as different research cohorts rely on different, sometimes not entirely compatible, definitions of memory traces in relation to speech. In psychological research, memory traces can be related to dynamic patterns of adaptation and the duration off the wash-out period (how long adaptations are maintained in the course of one or multiple sessions) in learning experiments. In neurosciences, memory traces tend to be biologically motivated and can refer to temporal profiles of responses in the brain or other specifics of the temporal dynamics of neuro-physiological processes.

[7] This effect seems to hold across modalities in automatized behaviors (Herman et al., 2009).

[8] Brain oscillations are repetitive electrical activities generated by neural tissue, usually in response to signals, or spontaneously.

The notion of time calls for a more dynamic, process-oriented perspective of representation. How are events represented in time? Which processes are involved in shaping the temporal dynamics of speech acoustics?

2.1.1 Processes involved in learning and using of speech representations

Probabilistic accounts of speech appear to rely on the 'ease of doing.' What do speakers 'do' when they speak? The 'doing' in articulation can refer to motor processes behind actively producing the actual speech gestures. The other part of 'doing' is involved in the cognitive processes involved. The ease of speaking can be related to the 'ease of processing' and 'ease of planning.' The processing and planning metaphors, again, tend to be couched in some notion of representation. It raises the question of what is being planned and what is being processed. It is not yet fully understood what is processed and what is planned in human articulation(Luo and Poeppel, 2007; Poeppel and Assaneo, 2020). More to the point, it is not yet fully understood how human behaviors (including vocalizing behaviors) are learned and represented in the human brain (Cao and Yamins, 2021; Cao, 2020). Given this, applying formal theories (which appear to rely on the existence of discrete events) to human behavior seems to lead to similar problems in the analysis of human and animal vocalizations. Namely, the problem of identifying appropriate levels of analysis in relation to the signal produced by the vocalizing individuals raised by Mercado and Perazio (2022):

> Given that it is not yet possible to experimentally identify the categories of units that are most salient from a humpback whales perspective, it is important to closely consider the acoustic variations that singing humpback whales produce and to take those characteristics fully into account when analyzing songs.

The focus on probabilistic models of language emphasizes the notion of predictability. This lead to models of speech production and speech comprehension as processes that involve 'planning' and 'prediction,' such that unpredictability or planning difficulties lead to articulatory variability (see, e.g., Buz et al., 2016). The planning metaphor exposes two challenging questions at once. The first challenging question involves identifying the differential contribution of planning and/or prediction to speech comprehension and speech production. The second challenging question seems to involve specifying the levels of description (the representations mentioned above) at which planning leads to efficiency and levels of description at which planning leads to error.

Planning supposes a top-down process that is, to some degree at least, controlled by the speaker. The actual behavior exhibited by the speaker is a highly automatized motor behavior[9]. Some automatized motor behaviors tend to not generalize across contexts and become more variable when people are asked to pay attention to what they are doing. For instance, asking people to attend to the tip of the pen they are writing with leads to a decrease in fluency and more variable performance in handwriting (Tucha and Lange, 2004). By contrast, the amplitude of fine-grain eye movements - saccade variability – decreases more when contextual variability is higher (which tends to focus attention, i.e., decrease attention to less variable dimensions) (Herman et al., 2009). That is, the extent to which people can control performance in fine-grain automatized behaviors seems to depend on the variability (and uncertainty) of the targets the planning involves. The automatized behaviors that benefit from attention are described as **skill** and appear to develop with help of constraints (uncertainty reduction) set by top-down, declarative knowledge[10] (Stanley and Krakauer, 2013). Automatized motor behaviors that do not benefit from attention (or manipulations to the variability of context) can be referred to as **habits** (Krakauer et al., 2006).

How do skills and habits relate to speech comprehension and speech production? The processes involved in skilled and habitual behavior can be simplified in terms of generalization learning, and a form of learning described as 'model-free.' Generalization learning describes a form of learning that is asymptotic and leads to a reduction of variable error over time, such that learned behaviors tend to converge on a 'model,' whereas 'model-free' learning tends to become increasingly divergent and fragmented (fine-grained) over time. In reinforcement-learning circles, it has been suggested that in contrast to model-based learning, where values are associated with rewards, values in model-free learning are associated with actions. This characterization, however, may be misguided. The form of learning that leads to an increase in the spectrum of observable behaviors that become increasingly

[9] Early evolved aspects of vocalizing responses can be viewed 'overlearned' motor behaviors. The communicative experience, at least in its initial stages, seems to be an integrated one. The whole body is involved. To be exact, the learner at the initial stages of her communicative development likely does not even know it has a body. A six-week-old infant has developed a relatively stable sleep-wake rhythm, and can control its gaze to follow the caregiver's face. It responds to prosody and produces vocal signals that become increasingly structured over time (Mampe et al., 2009; Wermke et al., 2021). It is yet incapable of recognizing that its hands as parts of its own body are different from the rest of the environment and can at times get very frustrated when its own hands move in an unexpected direction or fly in its own face. Meanwhile, its articulators become increasingly flexible and trained through feeding and the production of vocalizations. The vocal tract and the articulators are the first sites of controlled motor activity. This is important: at the time it utters its first words, a behavior that requires fine-grained motor coordination of breath, vocal tract, and articulators, the human infant is yet incapable of grasping a pea (or even a plum) with its index finger and its thumb. Talking is a complex, highly overlearned, skilled motor behavior. It is the most prevalent motor skill in humans and it is the first one we acquire.

[10] in which form declarative knowledge instantiates is not fully understood, it has been suggested that it may involve semantic representations. We will later, in chapter 5, suggest it that involves expectations and is implemented in the learning process through the latent structure of signals.

fine-grained within themselves, tends to be mediated by sequences of actions: behaviors conditioned on behaviors. The sequence decreases the variability of subsequent actions. Behaviors that occur in sequences (such as words in speech) tend to become increasingly fragmented (fine-grained) over time, while the variability at the transitions between the actions/behaviors seems to decrease (cf. Shmuelof et al., 2012; Shmuelof and Krakauer, 2014).

The discriminative learning paradigm suggests that both forms of learning can be explained in terms of discriminative learning and that the apparent differences between convergent and divergent forms of learning reflect the order in which signals are extracted from the structured environment by developing learners (Ramscar, 2019; Ramscar, 2021b; Ramscar, 2021a). In particular, the discriminative theory explains how **u-shaped learning** patterns are mediated by the structure of linguistic sequences (Ramscar et al., 2013e; Ramscar et al., 2018; Ramscar, 2019), and the protracted development of the prefrontal cortex in developing humans (Ramscar and Gitcho, 2007; Ramscar et al., 2013c). The u-shaped learning curve, in terms of discriminative learning, reflects an initial period of over-generalization (model-based learning), followed by a period in which behaviors (actions and signals mediated by the actions) become increasingly discriminated in relation to acquired models. That is, the discriminative theory introduces the idea that a decrease in variability (efficiency) is facilitated by learning from predictable, structured aspects of linguistic environments. This initial learning reduces the uncertainty, which in turn allows learners to make more fine-grained distinctions, which structure the uncertainty of subsequent experiences through constraints to expectations (see Ramscar, 2019; Ramscar, 2021b; Ramscar, 2021a, for summary).

This suggests that the grain at which articulations are resolved is facilitated by the order in which words occur in speech sequences. The uncertainty management mediated by the sequence ensures efficient transmission. It allows listeners to infer messages from noisy, 'incomplete' signals. Accordingly, there is ample evidence that those parts of signals that are inferable (very predictable in context) tend to be 'smudged,' although they are often still present in the acoustic signal (they are hypo-articulated) (e.g., Bell et al., 2003; Aylett and Turk, 2004). In speech research, this idea has become popular following work by Aylett and Turk (2004), which suggested that 'redundant' parts of sequences are reduced. Discriminative learning, by contrast, suggests that predictable parts of signals serve as cues in the initial phases of learning and, far from being redundant, may facilitate the structuring of sequences (Dye et al., 2018). The functional role of predictable aspects of signals changes with the speaker's experience. More predictable signals become increasingly uninformative cues to an increasing number of messages. The growing number of messages increases the functional load on sequences that facilitate the transmission. This increase in sequence length bears on structuring signals (i.e.,

regular aspects of signals, such as, for example, regular patterns of inflection). The functional roles of distinct aspects of the sequence change with the requirements of the communication.

All of the above suggests that the efficiency of communicative behaviors follows from learning and is facilitated by consistent signal structures. Moreover, predictable parts of signals, far from being redundant, can serve to mediate efficiency. This implies that information can only be defined in relation to the task (the context of communication). How is the task represented in the individual learner, and how will efficiency express itself under the constraints imposed by the task and a speaker's experience with it?

2.1.2 What are the functional requirements of human communication?

In linguistics, the communicative function is frequently quantified in terms of information that contrastive features contribute to the communicative process. In text- or abstraction-oriented analyses, contrastive features are assumed to be a set of discrete outcomes that approximate linguistic units: phonemes, morphemes, and words. Information is defined as some function of the (relative) probability of a communicative unit in context, $C_{blah} = \{w_1, w_2, w_3, ..., w_n\}$. A common metric in this kind of investigation is information content, or surprisal (Hale, 2001; Levy, 2008) a quantity that is often formalized in terms of a negative log probability of some event/unit (often a word) from a distribution of events/units defined by the context. Information content (or surprisal) is quantified as $-log(p(w_i))$. This metric supposes a context C_{blah} where all events probabilities add up to 1. This assumption is important, and it appears to be in conflict with the fact that human languages are learned and that C_{blah} is updated regularly so that we can expect the relative probabilities of words represented by C_{blah} to shift regularly[11].

> **The Alignment Problem:** Information is not a fixed quantity, it is an epistemic property; information is relative to a language user's experience (c.f. DeDeo, 2018; DeDeo et al., 2013).

The 'expectation-class' C_{blah} at varying points in speakers' development can take the following 'distributions' $\{0.4, 0.6\}$, $\{0.1, 0.2, 0.7\}$, but also $\{0.01, 0.05, 0.1, ..., 0.5\}$ and so on[12]. The uncertainty (entropy of the task/context) will change with experience,

[11] Further, because learning is discriminative, we can expect the category-structure at more fine-grained levels to be further subcategorized in course of learning.

[12] The problem of alignment of probabilistic expectations is essentially the problem of accounting for unattested quantities or taking the absence of evidence for 'evidence of absence.' The paradox

and the irregular distribution of words across contexts suggests that the probability mass function (the probability that a certain event will be associated with a given frequency, or recurrence rate) will change continuously with speaker experience. Given that both learning and maintenance of mutual expectation require relatively consistent sources of information, all of the above initially seems to suggest that information-theoretical quantities are not fit to measure information in human communicative systems. Yet information-theoretical quantities have been shown to be highly correlated to variation in speaker performance (and other aspects of behavior). Why? How is information in human communicative codes related to information in theory?

There are multiple parallels between human languages and the communication processes described by information theory: speakers' expectations seem to rely on statistical regularities, and the predictability of speech sequences at all levels of description seems to interact with sequence length (or duration)[13] (cf. Zipf, 1949). Moreover, similar to the communication process described by the coding theory (Hartley, 1928; Shannon, 1948), message transmission between humans is an incremental deductive process facilitated by uncertainty and discrimination. In contrast to speech (and language), Shannon's signal is produced by an invariant source, a so-called ergodic process, which guarantees that the statistical structure of the code is independent of how long the process is observed (sample size) or when the process has been observed (where the samples come from) [14]. The content of Shannon information is quantifiable because the uncertainty (entropy) of the context does not change in time, an observer can estimate the distribution from any given sample of reasonable size. **The definition of the process that underlies Shannon's definition of information is not compatible with the processes that lead to speech signal production and comprehension: learning to articulate intelligible signals and learning to discriminate intelligible from unintelligible signals in real-time.**

of assigning the probability 0 (or 1) to certain quantities is known as the Cromwell rule, after the English statesman Oliver Cromwell who, following the execution of Charles I in the summer of 1650, in a letter to the general assembly of the Church of Scotland, asked the Scots to reconsider their absolute certainty that Charles' son, Charles II, was to become the king of Scotland. The Cromwell rule in statistical modeling associates the practice of assigning probability zero to events and quantities with Cromwell's plea to the Scots: 'I beseech you, in the bowels of Christ, think it possible that you may be mistaken.'

[13] at least in context (see e.g. Piantadosi et al., 2011; Mahowald et al., 2013)

[14] **Ergodicity** is a property of a certain class of stochastic processes in signaling and dynamic systems. A process is ergodic if its statistical properties can be derived from a sufficiently long observation (a random sample of reasonable size). Any ensemble in the process (i.e., a collection of random samples/individuals at a particular point in time) represents the average statistical properties of the process. In other words, despite any variability in individual observations, both time and ensemble averages represent the process average. A process that changes erratically at an inconsistent rate is not ergodic. One example of a non-ergodic process is human vocal communication.

In an analysis of given name distributions, Ramscar (2019) observes that words mediated by context approach exponential distributions (which are memoryless and satisfy the conditions under which information-theoretical assumptions hold). Does this observation apply to speech signals too? Does it hold across languages? Languages, to varying degrees, use regularities in word order to manage the predictability of speech sequences. In addition, in speech, unlike in text and artificial communications envisioned by the information theory, code words (speech units) are time-varying quantities: the boundary of a speech 'unit' or speech 'chunk' is set by the rate at which speech signals (puffs of air modified by articulator configurations) unfold. The articulation rate is not constant. Words, syllables, and speech sounds represent predictable fluctuations in frequency, intensity, and the rate at which more or less predictable (informative) changes in the signal occur. Because speakers' expectations (about which part of the acoustic signal is predictable and which is not) vary with the context and speaker experience, which part of the transmitted quantities (represented by changes in the amplitude of different frequency modulation bands) can serve signals will also vary. Our initial analyses indicate that the transmissions of forms are regulated by context at multiple levels of description, including grammatical regularities (see Fig. 2.1). It follows from learning theory, that the extent to which speakers will be able to extract signals at more fine-grained levels of description (i.e., distinctions between semantic categories) will vary with experience. How is context implemented in speech?

More formal, 'implementational' notions of context, by contrast, rely on exemplar-based (essentially frequentist single-point forecasts[15]) quantities, represented by forms and form n-grams of varying length. The latter definition describes communication as an incremental process, where measurable quantities vary predictably in time (i.e., assuming a shared model[16] of expectations at each point in the sequence). The former definition implies an aggregate structured by a multitude of dynamical processes that vary in space (across individual speakers) and time[17].

Accordingly, the core ideas behind this work can be summarized as follows: In contrast to information-theoretic codes, human languages are learned. Learning at all levels of development appears to be bounded by constraints. Constraints will express themselves in speakers' ability to attend to different aspects of the

[15] while generative, or mixed, methods of modeling nested hierarchical structures explicitly commit to distributional (Bayesian) approaches, many of the simplifying assumption about the data-generating process implemented in these models collide with the initial purpose of a these approaches (c.f. Mcdonald et al., 2013); we will discuss this in the light of our findings and distributed, discriminative models in chapter 2

[16] aggregate distributions of underdetermined origin with hyperparameters that are uniformly distributed around the mean

[17] More to the point, the discriminative, systemic perspective exposes the usual problems involved in measuring time-varying quantities in complex systems/shaped by multiplicative processes/processes that work on multiple timescales (see e.g., Schrödinger, 1935).

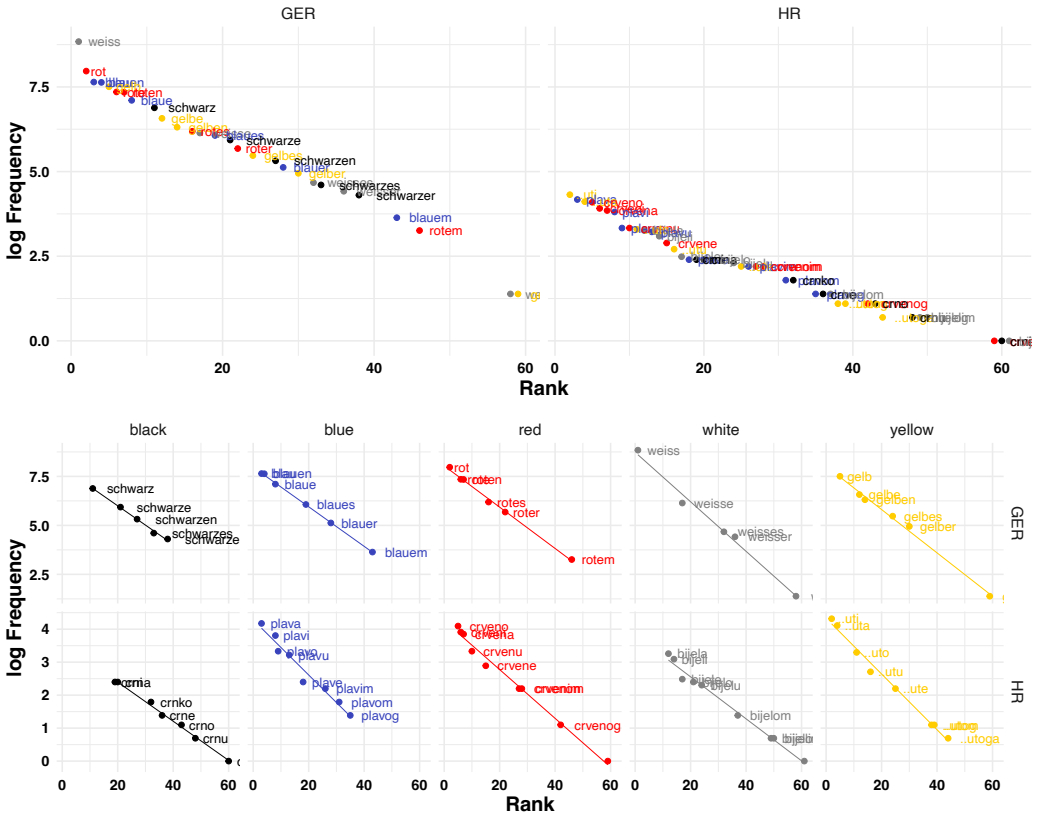

Fig. 2.1: Relationship between the log-transformed frequency and rank of Croatian and German color word forms from the CHILDES corpus. Bottom panel: morphological forms by language, sorted by color. The lines represent the linear model fit (a geometric distribution). Top: distributions of aggregated word forms.

input. Speakers' sensitivity to auditory contrast and different aspects of the linguistic structure is shaped by learning from structured input. Learning is discriminative so that speakers' prior experience changes speakers' present experience of the structure and the uncertainty related to traversing the structure in communication. The uncertainty associated with different functional aspects of the linguistic structure will express itself with drifts in the temporal resolution of speech sequences. Systematic changes in the temporal resolution will express themselves in the duration and quality of articulations. From this, it follows that the efficiency of speech signals will express itself as a systematic response to the uncertainty associated with the structural aspects of speech sequences and speakers' experience with them. This can, but must not necessarily entail shortening and compression, and as we will show, can lead to retention of functional contrasts that discriminate between structural patterns that are not necessarily expressed at the word and phoneme level and thus cannot be accounted for in terms of information provided by words and phonemes.

Returning to the relationship between form variation and functional load, global functional pressures on language tend not to be limited to phonetic categories.

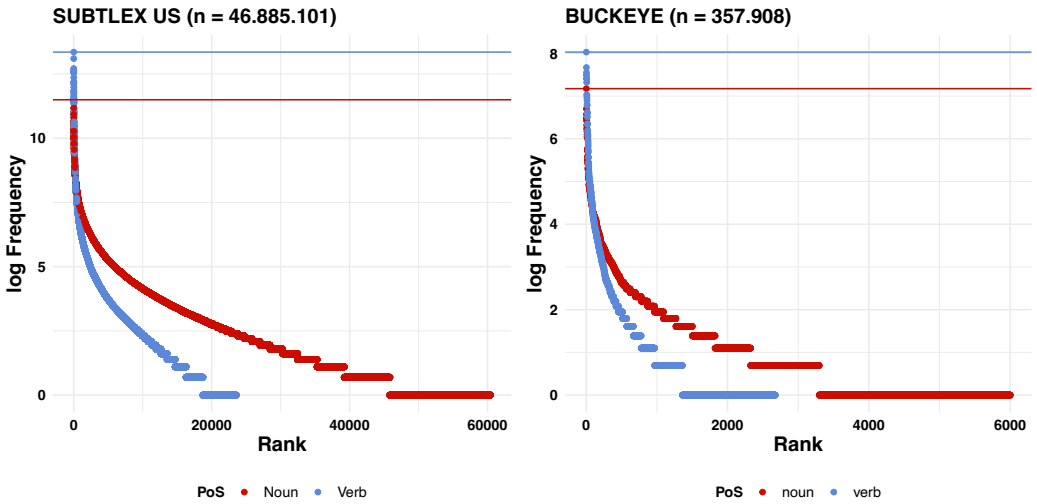

Fig. 2.2: Relationship between the log-transformed frequency and rank of verbs and nouns from transcribed speech (English.. The lines represent the frequency of the most frequent verb and noun in the corpus. The differences between the Buckeye Corpus of Spontaneous speech (left), and the much larger Subtlex corpus (right), are reflected in the asymmetric growth of the high-frequency head and the long-frequency tail of the distribution. In the left plot, the high-frequency head of the verb distribution, hosting 22 high-frequency verbs, makes up 44% of the probability mass. In nouns, the 22 most frequent nouns make only 12% of the probability mass. In the smaller Buckeye corpus, the difference in the high-frequency head of the distribution holds 28% of all verbs and 11% of the nouns. In the low frequency tail, 45% of all nouns and 49% of all verbs occur only once in the Buckeye corpus. In the Subtlex Corpus, 24% of unique nouns, 20% of unique verbs are found in the low frequency tail. In terms of the probability mass, the largest differences in how samples develop over time (as the size of the sample increases) seem to be reflected in the recurrence patterns of words from the high-frequency part of the distribution. We suggest that the differences in the head of the distribution capture differences in rates at which information is transmitted across individual speakers and registers.

Instead, functional pressures seem to operate at multiple levels of resolution such that their consequences may be represented at different timescales (they may initially go unnoticed and become evident after a delay). For example, developments that occur on slower timescales, such as organizational changes in civil societies (which, among other things, often involve the implementation of top-down policies and centralized systems of education that can affect the way people use words) seem to affect both the distribution of linguistic categories and the distribution of words these categories contain (Klingenstein et al., 2014; Ramscar, 2019).

Given that there is plenty of evidence that word realization co-varies with the uncertainty (i.e., variance) of the context, it is conceivable that the causal relationships (whether context co-varies with articulations or the other way around) work in both directions and that because **fine-grain local developments tend to be irregularly distributed in time (i.e., they are asymmetrically distributed in sequences)** they also tend to be invisible at the global scale (in aggregated speech data, for

example). More to the point, the **unlearning** of old habits and conventions reinforced by cultural technology takes more effort and time than learning of new habits and conventions in the context of existing ones, patterns of unconventional behavior, which tend to be implemented by individuals or small groups of individuals and occur in local burst. Because statistical models are conceptualized to perform well on aggregate data, i.e., models are typically designed to assume a certain type of distribution (or a certain type of distribution of distributions[18]), statistical models tend to be blind to accumulating evidence that occurs in bursts and/or has not been observed by the model (cf. Taleb, 2007; Taleb et al., 2020; Sampson, 2019). In other words, while statistical models of language typically rely on the unambiguous existence of linguistic representations and a stable global distribution, learning and language evolution appear to disrespect both the linguistic abstraction and the global distribution. The principles by which the evolution progresses are not yet not fully understood, and there is reason to believe that statistical models in the way they are currently implemented and applied are limited in their ability to answer the questions of development[19]. What is the alternative?

> **The Idea:** Regularly reoccurring words are periodic. They have stable ensemble averages[20] and vary in time. Fluctuations in the relative frequencies of regularly recurring words are informative. The variable error in the recurrence rates modulates speaker uncertainty about the rate at which **all** events unfold in sequences, and this leads to the reallocation of attention. The rate at which changes in the signals are extracted changes.

The latent structure of communicative distributions reduces the uncertainty about local variation (information rate). In other words, word co-occurrence statistics provide information about patterns of variation that violate the structure (represented by word co-occurrence statistics). In a similar way in which learning models deduce structure from patterns and detect signals by identifying criteria to discriminate signals that violate the patterns, local variation in signals reduces the uncertainty about the local probability of individual words and syntactic structures these words occur in. Given that context can be identified from patterns of systematic co-variance,

[18] which implicitly assumes some level of homogeneity in the aggregate (see Mcdonald et al., 2013, for discussion)

[19] In this vein, it is notable that current standard procedure in linguistic modeling involves excluding function words and low-frequency words (words that occur in the corpus less than 5 times) from the analysis. Function words make up 50% of the corpus probability mass. Depending on the corpus size, between 20-45% of unique types can be found in the low-frequency tail of the distribution. In other words, **most, if not all, current approaches to quantitative modeling of language cannot explain a substantial part of the data**.

[20] In statistical mechanics an **ensemble** is a number (usually a large one) of virtual copies of the system; in machine learning, an ensemble is a finite set of alternative models. Here we use the word ensemble averages to mean averages over a set of learners at a certain point in time.

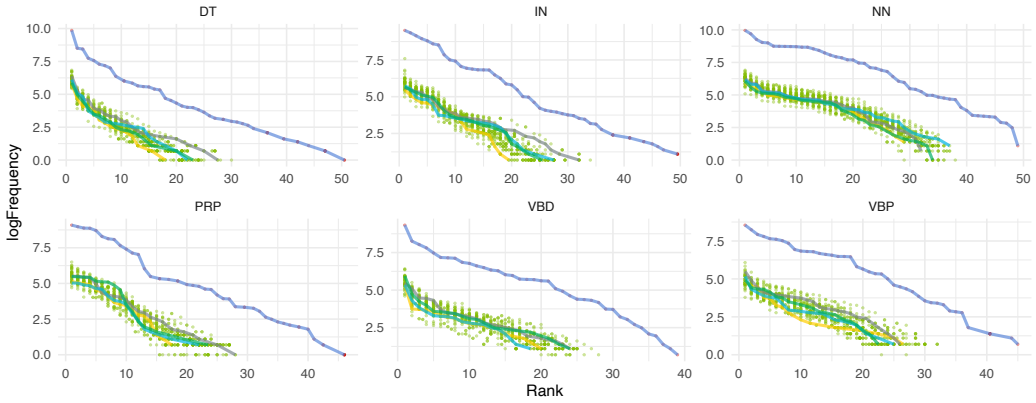

Fig. 2.3: The distribution of word-initial segments from spontaneous speech transcriptions in the Buckeye corpus. Apparently random variation in the articulation of word-initial segments from different part-of-speech categories (DT: determiners, IN: prepositions, NN: singular nouns, PRP: pronouns, VBD: past tense verbs, VBP: verbs, non-3rd person singular present) aggregates to a geometric distribution. Note that individual behaviors in isolation in many cases seem far from optimal (represented as variance in the scatter plots in green). The 'efficiency' appears to be achieved in the aggregate, the collective behavior.

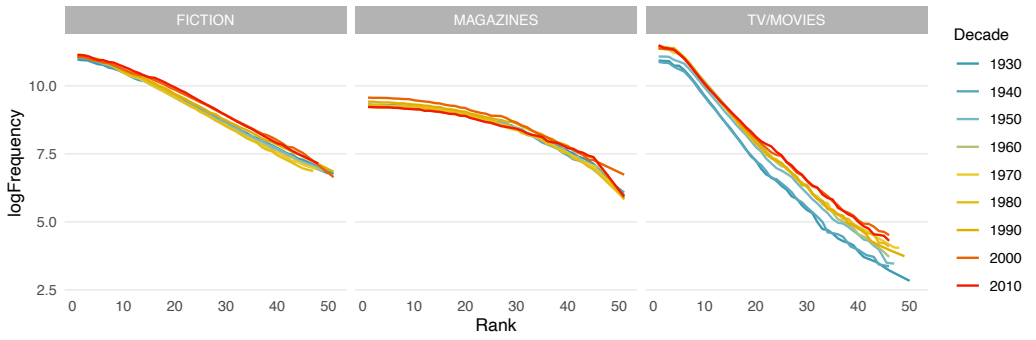

Fig. 2.4: **Distribution of utterance length in texts from novels, magazines, and subtitle data from 1930-2010**, Source: Corpus of Historical American English (COHA). Distributions slopes and shapes seem to reflect register differences, rather than corpus size differences, in aggregated text (120.000.000 fiction, 60.000.000 popular magazines) and aggregate speech transcripts (40.000.000 TV/Movie subtitles). The slopes of the distributions from both text corpora are more shallow, and the magazine corpus approaches the Yule distribution. Our analyses indicate that 'Yule'-like distribution are typically found where aggregation over items from closed categories occurs. In **'stratified' aggregates**, the rank-frequency relationship of individual types (or in this case lengths) vary across sub-samples obtained from different sources; this appears to be a scaling effect, and reveals itself as a **convex curve in the slope of the half-log plot**. The 'hump' in the middle part of the distribution, indicates that variable ranks are mapped on a fixed scale. The effect is well captured in the differences between length and frequency rank correlations in the 3 parts of the corpus (e.g., in 2010): FIC($r_(44) = 0.8450, t = 10.482, p < 0.0001$), MAG($r(44) = 0.6413, t = 5.5446, p < 0.0001$), TV($r(44) = 0.9709, t = 26.888, p < 0.0001$).

it appears that co-variance itself cannot serve as context outside of the higher-level context (because the covariance itself varies in ways that are bounded by the higher-

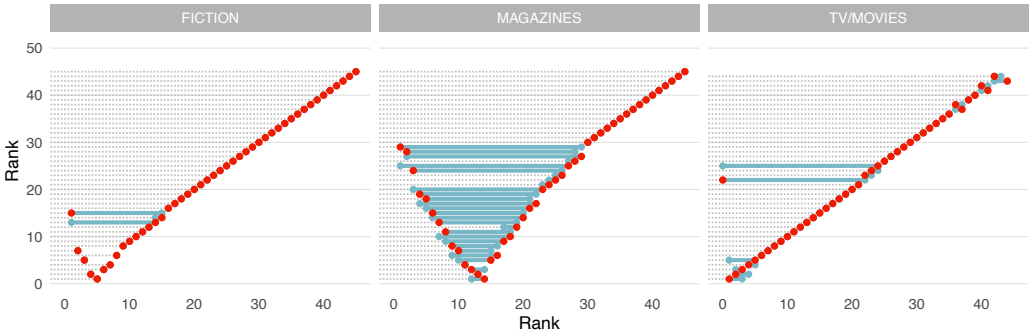

Fig. 2.5: Differences in the relationship between ranked frequencies and utterance length for decade 1950 and decade 2010, the blue lines represent the distance between utterance lengths at a given rank in the data collected between 1950-1960 and 2010-2020. Fewer blue lines indicate less 'rank-reallocation' in the corpus. While the frequency rank - utterance length relationship in novels and speech appears to be relatively stable across the decades, the values from the magazine corpus change considerably after 1970. We note here that novels and scripts of speech samples from movies and television series are typically written and edited by experts, and that editing of popular magazines before the introduction of computers and typewriters involved a meticulous, manual process (Open Culture, 2019).

level structure – local patterns of co-variation are associated with the local context and the messages).

More to the point, the information provided by the covariance between individual words seems to depend on the amount of uncertainty present at a certain point in time. Extracting information from word co-occurrence patterns requires that the context be identifiable first and identifying context requires knowledge about the distribution of contexts which seems to presuppose knowledge about the words. Or does it? How do we learn from exposure to continuous signals and how does this exposure shape our productions?

Learning increases speakers knowledge about the relationships between words and the world and words and other words. These developments affect the diversity of contexts words are associated with and the uncertainty of word clusters contexts provide support for. Given that words receive their communicative function in context, these systematic patterns of redistribution paint a relatively dynamic picture of form, function and structure, indicating that system states change as a function of learning. The implications of these findings raise questions: how can differences between structured sequences generations of speakers are exposed to in the initial stages of learning affect subsequent learning? How can the dynamics of the learning processes influence transmission?

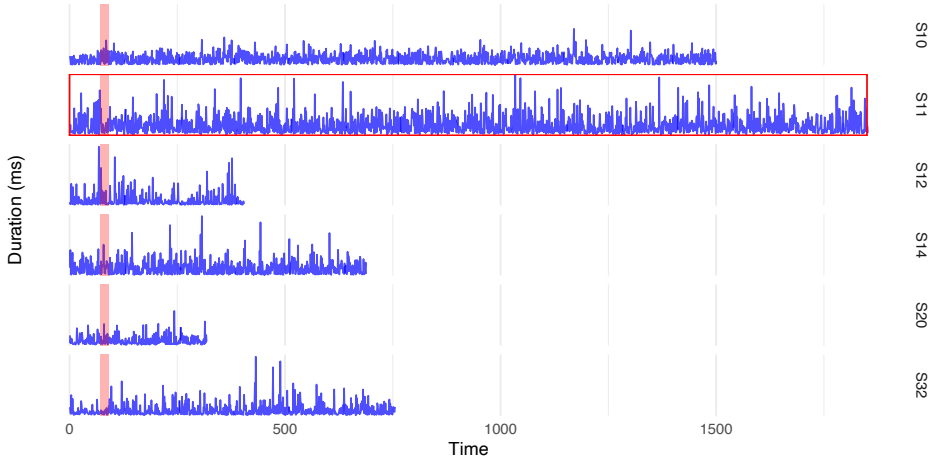

Fig. 2.6: Alignment in time as a 'systems solution' to problems introduced by learning: Ensemble averages vs. time averages in series of speech events: Shannon information requires that senders and receivers share their models of expectations at any given point in time. In human communication senders and receivers models of expectations differ. To extract information from signals speakers models of all possible acoustic events at a certain point in time must converge. It is likely both impossible and unnecessary for speakers models to converge on all possible acoustic events that can occur in time and across speakers. In other words, the solution to the learnability problem seems to involve alignment in codes, while the solution to the problems introduced by learnability seem to involve alignment in time.

Hypothesis: 'Reduction' and 'compression' can reflect temporal reorganization of speech sequences that is achieved through the optimization of the sequence structure: increase in the variability of sequence length, increase in the variability in local probability of high-frequency words or fillers, and adaptations to word order. The increase in the variability of the sequential ordering increases the dispersion of word level information and allows speakers to extract and articulate more fine-grained differences in the signal. This process is supported by learning to ignore uninformative aspects of the signal (periodic, overlearned dimensions of articulations).

Prior work on discriminative linguistics has mainly focused on text-based representations without explicitly addressing the constraints imposed by timings and time. The processing and production of spontaneous speech sequences, however, are subject to neuro-physiological and motor constraints in time. Both the rates at which informative changes in the signal can be processed and the rate at which informative changes can be articulated within a fixed timeframe are bounded by individual speakers' experiences. Motor learning involved in articulation is automatized so that at least some of the aspects involved in articulations will tend to converge relatively early in life. By contrast, the processes that lead to discrimination of acoustic signals adapt to the uncertainties of the contexts acoustic signals are represented in (Ami-

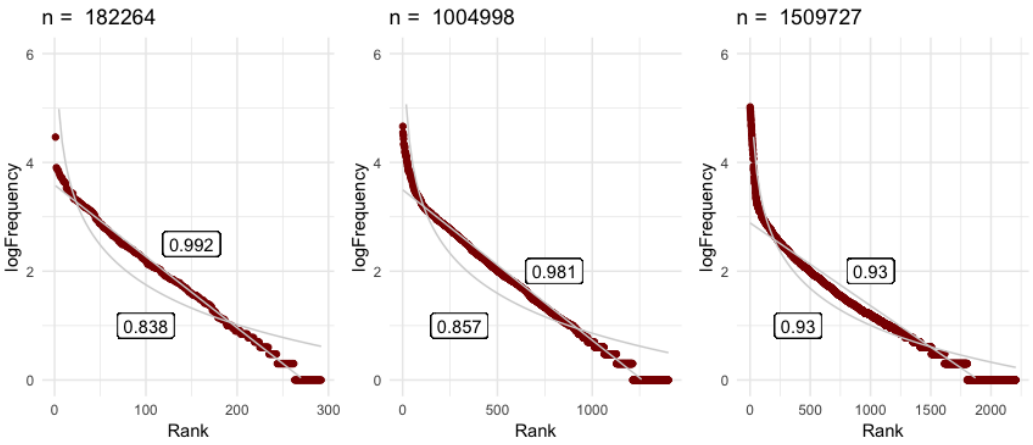

Fig. 2.7: Relationship between the log-transformed frequency and rank of number sequences from a sampling simulation at three subsequent steps (labels at the top of the plot show the size of the aggregate at each step). The labels show fits of the linear model for the relationship between log frequency and rank (geometric) and log frequency and log-rank (power law). The simulation implements a 'noisy copying' process by sampling from the original distribution and adding error to the copy. The error is added by multiplying the number by a fraction that varies with sample size. The simulation reveals that the major challenge of modeling the noisy process is in modeling a stable growth of the high-frequency head of the distribution (i.e., accounting for maintenance and moderate growth of the stable parts of the signal under noise). In communication, this part of the process is likely managed by systematic 'unlearning'.

tay et al., 2005; Amitay et al., 2006) and can be re-calibrated across the lifespan. The relationship between the temporal resolution of articulated gestures and the temporal resolution of signals will vary with speakers' experience. Because both the realization and the extraction of articulated forms depend on changes in signals that are predictable in time, it is questionable whether speech signal distributions can be analyzed meaningfully without accounting for time and experience-related distortions in time.

3 A Corpus Study: Speech form distributions in conversational English

Does systematic covariation in the usage patterns of word forms shape the way words and acoustic contrasts are articulated in conversational speech? This chapter addresses this question in terms of the discriminative theory of human communication. The theoretical framework explains how the distribution of events in communicative contexts shapes speaker expectations to maintain mutual predictability between language users. The analyses presented in this chapter are motivated by evidence that the distributions of words in the empirical contexts in which they are learned and used are geometric and provide sampling properties necessary for efficient communication and alignment of expectations. The analyses presented here show that the initial results extend to conversational speech. On data from a corpus of conversational English I show that the distribution of grammatical regularities and the sub-distributions of tokens discriminated by them are also geometric. Further analyses reveal a range of structural differences in the distribution of types in parts of speech categories that further support the suggestion that linguistic distributions (and codes) are subcategorized by context at multiple levels of abstraction.

3.1 Form distributions in human communication

Words produced in conversational speech often differ substantially from the acoustic signals supposed by canonical dictionary forms (Port and Leary, 2005; Ramscar and Port, 2016). The extent to which articulated signals deviate from dictionary models is correlated to average word frequency, such that there is a general tendency for shorter and faster articulation in more frequent words. This property of speech codes is often taken to suggest that human speech is shaped by the competing requirements of maximizing the success of message transmission while minimizing production effort in ways similar to those described by information coding solutions for electronic message transmission. There are, however, some critical differences between speech and the communication model described by information theory: whereas information theory is concerned with defining the properties of variable length codes optimized for efficient communication in discrete, memoryless systems,

human communication codes, at first blush at least, appear neither systematic (Ramscar, 2019) nor systematically discrete (Arnon and Ramscar, 2012; Ramscar and Port, 2016) or memoryless (Ramscar et al., 2014).

With respect to the first point, **systematicity**, it is relevant that humans, unlike transistors projected by the information theory, learn to communicate by gradually discriminating functional (task-relevant) signal dimensions from samples to which they are exposed. This property of human communicative codes (that they are learned, and not shared from the onset) is important, because humans learn from samples that are diverse. Lexical diversity in language samples increases non-linearly in space and time, and the divergence between the samples to which individuals are exposed increases with exposure [1]. A system defined by a probabilistic structure would seem to require that events are distributed in a way that allows the relationships between events probabilities to remain stable independent of the sample size, yet the way that words are distributed across language samples suggests that human languages do not satisfy this requirement.

Considering the second point, **discreteness**, although writing conventions lead to some systematic agreements about what linguistic units are, such that words are often thought of as standard discrete linguistic units, speech appears to be different. Human intuitions on what constitutes a unit boundary in speech tend to diverge as exposure increases. When literate adults, non-literate adults and children are asked to divide the speech sequence into units, their intuitions on where any given sequence should be split into multiple units exhibit a systematic lack of agreement (Scribner and Cole, 2013); similar effects have been observed when people are asked to discriminate phonetic contrasts (Munson et al., 2010). Moreover, early speech segmentation strategies were shown to vary with the native language (Nazzi et al., 2006) and vocabulary development (Kidd et al., 2018).

As for the **memorylessness**, which supposes a distribution of events such that an event's probability is independent of the way it is sampled, it has been shown that increased exposure to language leads to a decrease in the informativeness of high-frequency tokens relative to words they co-occur with, such that the informativity relationships between words appear to be unstable across cohorts (Ramscar et al.,

[1] Specialized contexts require specialized vocabularies: while *distribution* is a relatively frequent and highly functional part of the authors active vocabulary, the frequency at which it occurs in speech across all speakers and contexts is much lower. This property of word distribution is reflected in the fact that words are not equally probable across registers: 1 million words of the Corpus of Contemporary American English (COCA) (Davies, 2010) on average contain 106.62 tokens of type distribution in the science register, compared to 8.97 tokens in the spoken register and 3.64 tokens in the fiction register. The bursty distribution of low-frequency words across communicative contexts bears out that language samples speakers are exposed to and learn to use in their daily communications diverge considerably in the low-frequency range (Ramscar, 2019). As a result of this, average counts in low-frequency types seem to not allow a systematic description of predictability.

2014). For instance, the information the word *blue* provides in isolation and in relation to words that tend to follow up on it, changes systematically as people successively hear about *blue skies*, *blue eyes* and *blue berries* (but not *blue bananas*) etc. at different rates. This effect of lexical competition on speakers' expectations increases non-linearly with the number of blue covariates that speakers encounter.

To summarize these points, it seems clear that adult expectations about events and their probabilities vary with experience and the context in which they occur. This in turn seems to suggest that the increasing divergence between individual speakers' models will lead to an increase in communication problems between speakers. Nevertheless sufficiently successful communication between speakers of different experience levels is not only possible, but relatively common. How?

Recent work by Ramscar (2019) addresses these apparent communication problems from the perspective of discriminative learning, and suggests that unlike the predefined source codes in artificial communication, human communicative codes are subcategorized by systematic patterns of variation in the way words and arguments are employed. The empirical distributions discriminated by these patterns of variation both serve to minimize communicative uncertainty and make unattested word forms predictable in context, thereby overcoming some of the problems that arise from the way that linguistic codes are sampled. In support of this argument, Ramscar presents evidence that the empirical distributions shaped by communicative contexts are geometric, and suggests that the power laws that commonly characterize word token distributions are not in themselves functional, but rather result from the aggregation of multiple functionally distinct communicative distributions (Newman, 2005). Importantly, unlike power law distributions, the geometric distributions are sampling invariant and thus directly satisfy many of the constraints defined by information theory (Shannon, 1948; Hartley, 1928). Moreover, geometric distributions also appear to maximize the likelihood that, independent of exposure, learning will lead to speakers acquiring similar models of the distributions of communicative contrasts in context. This in turn ought to allow for some degree of mutual predictability, which would help explain why human communicative codes actually work as well as they do.

An interesting finding with respect to the distributional structure of communicative codes comes from an analysis of names (a universal feature of communication that is often ignored in grammatical theories), and in particular the distributions of English and Korean first names (Ramscar, 2019). Analyses of given name distributions show that historically, prior to the imposition of constraints (name laws) to name sequences, first name distributions across a range of cultures and populations had nearly identical geometric distributions. Names are a unique aspect of language in the way they are used tends to be highly regulated in modern states administration.

Functionally, name sequences serve to discriminate between individuals, and thus it follows that imposing constraints on the way name sequences are used and fixing the distributions of name tokens by law will affect the discriminatory power of those distributions. The 20th century is characterized by large global increases in population sizes; that is, the number of individuals that name distributions discriminate between has increased noticeably. In western populations this has had two consequences: first, the administrative constraints on last name distributions seem to have lead to an increase in functional load on the distribution of the first names in direct proportion to increases in population. Second, this seems to have led to an increase in the diversity of regional first name distributions across very large countries such as the United States, where first name distributions follow power laws. An interesting aspect of the aggregate distribution, is that although first name distributions in the US as a whole follow a power law, the distribution of names in the individual states partly still show better fits to the geometric, indicating that the evolving shape of the aggregate distribution may reflect shifts in the uncertainty of diversifying local distributions (Ramscar, 2019; Newman, 2005).

These results suggest that across space and time, discriminative codes somehow respond to the various communicative pressures imposed by the environment in ways that sustain the sampling invariance that seems to be crucial to efficient, systematic communication. This is interesting, because individual contributions to the name pool appear, at least intuitively, to be somewhat random. These findings offer some interesting perspectives on the apparent similarities and differences between communication in the human and information theoretical sense, and raise some interesting questions in regard to speech. To what extent are other aspects of speech codes shaped similarly by the competing pressures of providing sufficient contrast to communicate about necessary, or mentionable distinctions, while at the same time retaining a sufficiently stable structure to allow speakers to be mutually predictable as they learn about these distinctions? Is the variance in the way people articulate speech signals a consequence of the uncertainty of the context in which they are learned and used, and does this variance have a communicative function?

The following sections briefly review the theoretical background to the present analysis. Section 3.1.1 reviews some key findings about linguistic distributions that appear to support their communicative function. Section 3.1.2 describes some of the implications these findings for speech and finally, section 3.1.3 lays out a set of explicit predictions derived from this theoretical analysis. These are then examined in the rest of the chapter.

3.1.1 Grammar as Context - Convention Shapes Learning Shapes Context

It seems that human communication codes are not shared in the predefined way that information theory supposes (Ramscar, 2019). Natural languages are learned from exposure to specific, incomplete samples, and these can diverge considerably across individuals and cohorts. The bursty/uneven distributions of low-frequency types observable in large language samples indicate that a large portion of these types will be either over- or underrepresented across the communicative contexts any individual speaker is exposed to. This in turn suggests that any communicative system operating on global word probabilities will be inefficient and unsystematic. At the same time, the fact that regularities in human languages can be consistently captured and shared through linguistic abstractions at different levels of description suggests that speech codes provide speakers (and learners) with probabilistic structures that are sufficiently stable to ensure that most important linguistic conventions will be learnable from samples all speakers are exposed to. For example, Blevins et al. (2017) suggest that the existence of regularities in the distribution of morphological forms serves to offset many of the problems that arise from the highly skewed distribution of linguistic codes, since the neighborhood support provided by morphological distributions makes forms that are otherwise unlikely to be attested to many speakers inferable from a partial sample of a code.

Languages are learned from the input and what is learned is contingent on the variety of cues present in the input, the temporal relations between the cues, and the order in which the events unfold in time (Ramscar and Yarlett, 2007; Ramscar et al., 2010; Arnon and Ramscar, 2012). Importantly, a critical property of learning is that speaker expectations are biased by prior experience (Ellis and Sagarra, 2010; Ramscar and Port, 2016; Nixon, 2020). That is, experience constrains which signal dimension is perceived as variable (and informative) and by consequence, the signal dimensions they have learned to assume invariant (and thus not informative). From this perspective, context can be described in terms of a sufficiently invariant discriminative structure (a signal dimension that does not change significantly with experience) acquired through learning. These predictable signal dimensions serve to facilitate learning of underlying discriminative patterns of fine-grained variation through uncertainty management.

Speakers' ability to interpret pseudowords in context (McDonald and Ramscar, 2001), for example, *He drank the* $dord$ *in one gulp* offers another illustration of this point. Here the lexical context provides sufficient support for the inference that *dord* is likely a drink of some sort, regardless whether it is familiar to the speaker, or correlated to a real life experience or not. (In the former case, if *dord* were to occur more

regularly and in correlation to an actual bottled or cupped substance in the world, it would become a part of the vocabulary, losing its non-word status.) These kinds of context effects appear to rely on the fact that in sequences *drink milk*, *drink water* and *drink beer*, *drink* systematically correlates with words that in turn co-vary with the consumption of fluids, unlike *eat apple*, *eat banana* and *eat chicken*.

Once this higher level abstraction, or *context*, of two distinct types of consumption is established, it allows the speaker to predict from *drink* that the upcoming part is probably not solid and discriminate from the signal the specifics of the drinkable non-solid experience, *milk*, *water*, *beer* or *dord*, eventually. This emphasizes how once the ability to distinguish between *contexts* is sufficiently learned (and invariant), it renders the immediate sensory experience less relevant and allows the speaker to abstract from the experienced. What is talked about must not be seen or felt to be believed or be subject to experiential constraints. Concrete properties can be assigned to objects which are not present, or part of the experienced world and in fact may not even exist (yet). In other words, we can infer that *drink*, *chicken* and *dord* mean different things as a function of the lexical contexts they tend to not appear within, without them necessarily being associated with a real-world experience.

Given the discriminative nature of learning, it follows that exposure to language samples containing this kind of systematic co-variance structure will lead to the extraction of clusters (subcategories) of items that are less discriminated from other items that occur in the same contexts than they are to unrelated items. Further, there is an abundance of evidence that patterns of systematic co-variance of this kind provide a great deal of information, not only at lexical level (where semantically similar words typically share co-variance patterns), but also at a grammatical level (Ramscar, 2019). For example, in English, different subcategories of verbs can be discriminated from the extent to which they share argument structure with other verbs. The way that verbs co-occur with their arguments appears to provide a level of systematic co-variance that nouns appear to lack (Levin, 1995). For instance the following sentences would be considered grammatical:

1. John *murdered* Mary's husband.
2. John *ate* Mary's husband.
3. John *chewed* Mary's husband.

Whereas the following sentence would not:

4. John *ran* Mary's husband. (*)

One reason for this difference is that *chew*, *eat* and *murder* share a similar pattern of argument structures (co-vary systematically) in a way that *run* does not. By contrast,

the kinds of grammatical contexts which predict nouns (noun phrases) appear to allow any noun - the sentence is grammatical - irrespective of its likelihood (although, obviously, these will vary widely according to context).

5. John *ate*.
6. John *ate* cheese.
7. John *ate* cheese slowly with a toothbrush.

In other words, the systematic co-variance of verbs in their argument structures appears to constrain their distribution in context far more than is the case for nouns.

8. Mary *loved*. (*)
9. Mary *loved* cheese.
10. Mary *loved* cheese slowly with a toothbrush. (*)

Accordingly, the distributional patterning of verbs thus appears to reduce uncertainty not only about the lexical properties of upcoming parts of a message, but also about the messages structure. Or, in other words, because verbs take arguments, there ought to be less variance in their patterns of co-variation, and this ought to lead to less overall uncertainty in the context of verb arguments. Consistent with this, Seifart et al. (2018) report that slower articulations and more disfluencies precede nouns then verbs across languages, raising further questions about the kind of information that is communicated by variational patterns in speech, and in particular whether, and to what degree, this kind of sub-lexical variance actually serves a communicative function.

In the next section we review some evidence that suggests the interactions observed between uncertainty and articulatory variation may indeed be functional.

3.1.2 Sublexical variation in context

It is relatively established that isolated word snippets extracted from connected speech tend to be suprisingly unintelligible when presented outside of their original context. By contrast, when these reduced variants are presented to speakers in contexts in which they were produced, speakers are able to identify the word without difficulty and often even report hearing the full form (Ernestus et al., 2002). This indicates that spontaneous speech forms vary in ways that are informative in relation to the context they are articulated in, and that frequency effects might be an epiphenomenon of a process bootstrapped by contextual diversity (instead of word frequency). Consistent with this, the effect of frequency on variation in speech form production has been shown inconsistent across registers, speakers, lexical classes

and utterance positions and there are opaque interactions between context, lexical class and frequency range.

At first, these inconsistencies could appear to limit the scope of functional theories of speech sound variance. At closer look however, they are also informative: to date the effects that are stable enough to be taken as evidence for functional theories are mostly to be found in content words from the mid-frequency range. The effects observed in the remaining (by token count significantly larger) parts of the distribution, which constitutes both very frequent and very rare words, tend to not align with those in words from the mid-frequency range. For example, while function words, high frequency discourse markers and words at utterance boundaries account for the largest portion of variance in speech and probability mass in a corpus, their exclusion from the analysis is such a common practice that it might be considered a de facto standard (Wedel, 2012). Against this background, it is noteworthy that Bell et al. (2009) report a divergence in the extent to which the articulation of function and content words across frequency ranges is affected by both frequency and the conditional probability of the collocates. While duration in content words is well predicted by the information provided by the following word, but not the preceding word, the effect decreases as the frequency increases and shows a reverse pattern in function words. Similarly, Van Son, Pols, et al. (2003a) report a reversal in the correlation between reduction and segmental information in low-information segments and segments at utterance boundaries. The effect of information content [2] is reported to be limited by a *hard floor* in high-frequency segments; that is, both most frequent segments and words fail to support the hypothesis that speech form variation is a matter of probability. This might be taken to suggest that standardizing the exclusion of misfits is actually a little controversial, given that they seem to outnumber the units (be it segments or words) which are typically taken to confirm the hypothesis, and given that the segments and words from this frequency register also seem to account for the largest part of variance in speech (see also Sampson, 2013; Popper, 2005).

At the surface, it seems that frequency effects, which are correlated to contextual diversity [3], might fail to explain variation exactly at those points where contextual diversity tends to diverge from frequency – namely in the tails of the distribution. What is it about the words from the tails that makes them so different? Perhaps it has something to do with the distribution of uncertainty across word categories that the tails typically host?

[2] an information measure sensitive to the relative frequency of segments and words in context in relation to their average probability in the corpus

[3] because more frequent words are more likely to appear in more contexts

As noted in this chapters introduction, low frequency words typically occur in specific contexts only, and their relative predictability is regulated not only by the argument frame (surrounding words), but typically also the register that allows for the bandwidth/rate that the transmission of specific information requires. For example, talking about *register specific transmission rates* requires that the speakers know that relative probabilities of communicative events and the amount of new information they can serve to communicate are determined by the target audience and the communicative purpose (what the average target doesn't know yet). Therefore, low-frequency words and messages, however well embedded in the argument frame they might be, are noise to targets alien to the register (i.e. toddlers and all other people not interested in the probabilistic structure of speech codes). High frequency words on the other hand tend to be dispersed across many lexical contexts in which they achieve their communicative function. The most frequent words, function words, achieve their function by providing context to other, less frequent words. That is, function words, high-frequency discourse markers and fillers seem to be at a different pole of a gradient function (defined by the lexical context on the one side and lexical contrast on the other). These different functions provide information that distinguishes between two different kinds of uncertainty speakers possess: how signals are structured (form, which relies on shared expectations) and the specific signals they serve to transmit (function, which relies on differences in expectations).

There are several clues that indicate that speech form variation provides information about uncertainty (which is a dynamic property of the distribution), rather than the segment or the word itself. Observations of *deviant articulations* at various levels of description suggest that seemingly random and noisy variation in the speech signal is correlated with the uncertainty about the upcoming part of the message. As an example, vowel duration in low- and mid-frequency content words is correlated to the information provided by the upcoming word (Bell et al., 2009). Words in less predictable grammatical contexts are on average longer and more disfluent (Tily et al., 2009). These fluctuations in duration and sequence structure have been shown to inform listeners' responses. For instance, the duration of common segments in word stems differ between singular and plural forms (Salverda et al., 2003). Speakers appear to use acoustic differences in word stem as a cue to grammatical context (plural suffix) and incongruence between segmental and durational cues lead to delayed responses in both grammatical number and lexical decision tasks (Kemps et al., 2005). Similar effects occur at many other levels of description, for example, disfluent instructions (*look at the uhm camel*) lead to more fixations to objects not predicted by the discourse context (Arnold et al., 2003) and facilitate prediction of unfamiliar objects (Arnold et al., 2007). Taken together, the variation in those part of the speech signal that are typically not available in text corpus analysis, seems to contribute to the coordination of speaker expectations.

The uncertainty signaling, in turn, seems to enhance speech perception. The occurrence of silent and filled pauses has been shown to contribute to the perception of fluency (Bosker et al., 2013) and intelligibility (Bosker et al., 2014b) as well as improved recall (Fraundorf and Watson, 2011; Diachek and Brown-Schmidt, 2022). Importantly, however, neither artificially slowed-down speech samples nor samples modified by insertion of pauses are then perceived to be more fluent or intelligible. Instead, in both cases these manipulations have been shown to result in impaired performance (Cooke et al., 2014). Accordingly, the fact that listeners easily interpret reduced sequences from context and reject speech artificially altered to mimic completeness and fluency indicates that hearers are highly sensitive to violations of their expectations about how natural speech should sound, and not that they have a preference for completeness and slow and extreme articulation. Yet despite the evidence that sub-lexical variation shapes speaker expectations about the upcoming content, its contribution to successful communication as an informative part of the signal has remained relatively unexplored to date.

However, it is clear that any quantification of the communicative contributions of sub-lexical variations in context will depend on a consistent definition of context. That is, in order to address the extent to which the quality of articulation and the observed variance in the signal interact with the remaining uncertainty about the message in general terms, it is necessary to first formalize a consistent subset of higher-level abstractions that systematically co-vary in the degree to which they contribute to uncertainty reduction. The contrast between these subsets can then allow these effects to be analyzed independent of the specific context of any given utterance.

So far context has been formalized as corpus document (Wedel et al., 2018; Jones et al., 2017), local lexical environment (Bell et al., 2009; Wedel et al., 2018; Piantadosi et al., 2012) and syntactic context (Wedel et al., 2013a; Tily et al., 2009). As abstraction frames in quantitative research are limited by technology, most of the current approaches rely on extrapolation of sequences uniform in size, which appears to lead back to the problems of systematicity, discreteness and memorylessness in representations. The distribution of information across lexical n-grams, words and speech sounds varies across speaker, and varies across languages. For instance, Piantadosi et al. (2012) report that a words probability in its lexical context (defined as n-grams of length 3-5 tokens) is a better predictor of average word length than frequency. Notably, a closer look at the results shows that **the extent to which the lexical context and n-gram length explain the effect varies across different languages.** The size of the effect appears to reflect the morphological complexity of the language, which determines the degree to which arguments are supported by other elements in the frame (e.g. noun endings in German case). This does not mean that word probability in context is not correlated to information, it simply

underlines the fact that languages differ in the extent to which they rely on regularities in the relationships between words (which English does more than Korean) instead on regularities in the relationships between morphological forms (which Korean does more than English).

The present study aims to examine these differences by comparing the distribution of speech form variants from a corpus of conversational English and Korean in the context of morpho-syntactic regularities.

3.1.3 The Present Study

In comparison to written language, speech often appears to be messy. Instead of the well-formed word sequences that characterize text, spontaneous speech sequences are typically interrupted by silent and filled pauses, left unfinished, depart from word-order conventions, frequently miss word segments or whole words and rely on clarifying feedback which tends to be short and grammatically incomplete. In consequence, the token distributions that underlie the information structure of written and spoken language differ substantially.

For instance, nouns are less lexically diverse in spoken English then in writing[4], whereas English adjectives tend to be more lexically diverse in speech. While reading and writing are self-paced, speech gives both speakers and hearers less control over timing. This suggests that that the moment to moment uncertainty experienced in communication may differ in speech as compared to written language, and it may be that more effort is invested in uncertainty reduction in spoken than in written language. From this perspective the increase in the lexical variety in prenominal adjectives, which in English reduce uncertainty about upcoming nouns (Dye et al., 2018), might be functional in that it may help manage the extra uncertainty in spoken communication, hence raises the question of the degree to which these and other variational changes in spoken English are indeed informative and systematic.

These considerations also suggest that the results of previous analyses of the distributional structure of lexical variety in communicative contexts conducted on text corpora can only offer indirect support when it comes to answering to questions about the communicative properties of speech. To address this shortcoming, we conducted a corpus analysis of conversational English (Pitt et al., 2005) to explore the extent to which the distribution and the underlying structure of the grammatical context words are embedded in interacts with speech signal variation observed across lexical categories. The goal of this analysis was explore the structural

[4] based on measures derived from the Corpus of Contemporary American English (COCA)

properties of grammatical regularities in speech, and their effect on the distributions of the lexical and sublexical contrasts that they discriminate between.

The analysis was conducted in two stages. Part one, presented in this chapter addresses the distribution of grammatical and lexical contrast in speech and aims to answer the following questions:

- Are distributions of grammatical regularities in speech sampling invariant?
- How do recurrence patterns of grammatical categories and speech sequences inform learning?
- Are distributions of subcategorization frames and types they distinguish between geometric?

Part two of our analysis, presented in chapter 4 assesses the concrete consequences of the sub-lexical variation observed in the speech signal, and relates these to the results presented in section 3.3, addressing the following questions:

- Are the inconsistent effects of frequency on speech sound variation across categories correlated with structural and distributional aspects of the grammatical and lexical contexts they populate?
- Finally, and perhaps most importantly, is the resulting sublexical variance systematic?

3.2 Materials and methods

3.2.1 Data

The Buckeye Corpus (Pitt et al., 2005) contains phonetically transcribed speech from informal interviews with 40 speakers from Columbus, Ohio. The $286,982$ words are annotated with a set of 41 standard aligner phone labels expanded by a set of markers for manner of articulation (nazalization, flaps, glottal stops, and retroflex vocalization). The corpus version we used was extended by Dilts (2013) with measures of segmental deletion; dictionary form alignment; and deviation rate normalized by word length, speech rate, and backward and forward conditional probabilities of word ngrams. For the analysis reported here, we excluded from the corpus 8426 words with missing or incomplete duration variables. The data set and the code for the analysis can be found at `https://osf.io/bqepj/`.

In each of the 1-hour interviews, the 40 speakers (who are balanced by age and gender) showed an enormous amount of variability (as assessed by phonetic transcription) in the speech signal. Overall, only 40% of the words are produced in their citation form. Only 38% of word types tend to appear in their non-citation variants more often and the propensity of individual speakers to pronounce word types in their citation form varies widely (between 36% word types and 67% word types). The word *that* appears in 313 variants, including *d ah tq*, *m ah t*, *z eh tq*, and *ng ah*.

We extracted for each citation form and parts-of-speech combination the number of variants observed in the corpus by citation form. The relative frequency counts for each form by parts-of-speech label were taken from the spoken part of COCA, an 80 million token subcorpus of Contemporary American English from transcripts of unscripted conversation on TV and radio programs.

3.2.2 Analysis of Probability Distributions

Plotting a frequency distribution on a log-log plane, with log frequency on the y axis and log of rank order on the x axis, is a common method in the analysis of probabilistic structure. A linear plot indicates that the data conforms to Zipf's law because Zipf's law assumes an exponential increase in the time rate (rank). That is, a linear plot confirms a power law, while distributions we observe here and other variants of aggregate distributions (e.g., Zipf-Mandelbrot) are reported as an anomaly. The latter usually entails the introduction of additional parameters to fit the distribution back to power law.

Because linguists have so far only searched for power laws, the distributions we observe here are, when found, reported as an exception (Arbesman et al., 2010). Ramscar (2019) argues that empirical linguistic distributions ought not to be expected to follow power laws. Rather, because learning and mutual predictability require a regular distribution of events over time, human communicative codes ought to be expected to have distributions that retain their structures over time. Accordingly, following Ramscar (2019), we employed log-linear plots in these analyses. That is, the linear decrease in probability over discrete time defines a time invariant communicative distribution while the exponential decrease in probability does not. To asses the extent to which the method captures this property, we apply it to a set of subsamples drawn from the original data.

Figure 3.1a,b shows results from a simulation study capturing fits of the analyzed categories to a geometric distribution and a power-law distribution, respectively, over the first 2500 words from each of the 40 speaker subsamples. The two bottom

row panels show the fits to geometric (Figure 3.1c) and power law (Figure 3.1d) across 40 random subsamples varying in size between 652 and $19,363$ tokens. As we can see in Figure 3.1, fits to power law vary with sample size and source across all categories. In contrast, fits to geometric remain relatively stable in empirical distributions independent of sample source and size. This is not the case for aggregate distributions. Accordingly, this method appears to capture the critical property of communicative distributions addressed in this paper.

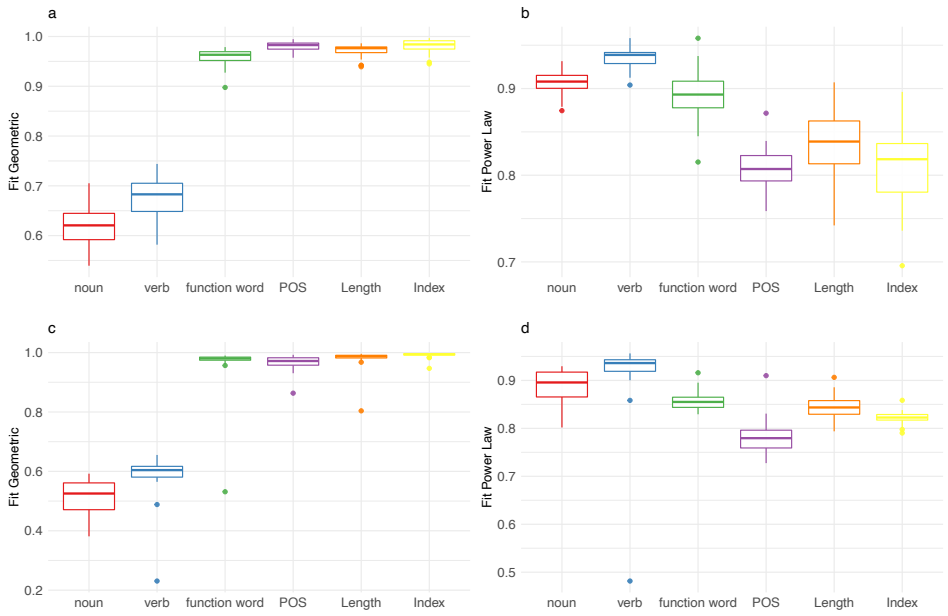

Fig. 3.1: Boxplots of fits to geometric distribution (**a**,**c**) and power law distribution (**b**,**d**) for categories analyzed in Sections 3.3 and 3.4 for the first 2500 words by 40 speakers (**a**,**b**) for 40 random samples ranging in sizes between 652–19,363 (**c**,**d**).

3.2.3 Statistical Analysis

The results presented in Chapter 4 were analyzed with a generalized additive mixed-effects model (GAMM) (Wood, 2006; Hastie and Tibshirani, 1990), working with the *mgcv* package for R. GAMMs are used for the analysis of complex, often nonlinear patterns involving the interaction of two or more numeric and factorial predictors. Instead of using polynomial functions, GAMMs introduce smoothing splines. A smoothing spline with one predictor fits a curve over multiple basis functions. Smoothing splines with multiple predictors fit multidimensional surfaces. These features allow us to explore the data structure and interactions between frequency range, context, and lexical category and to reduce the model complexity by identifying relevant dimensions which eventually allow us to identify linear effects (see chapter **??** for discussion).

3.3 Part-Of-Speech Token Distributions – Why Parts of Speech?

It is clear that many important regularities in human languages are consistently captured by high-level linguistic abstractions such as, for example, parts-of-speech categories, indicating that languages may be sufficiently structured to allow the discrimination of various functional parts of codes at various levels of abstraction. Ramscar (2019) suggests that the probabilistic co-occurrence patterns of words and phrases serve to discriminate subcategories of signals (and hence codes) and that, as well as serving different communicative purposes, these subcategories form distributions that facilitate speaker alignment at various levels of analysis. This raises an obvious question: do the distributional properties of structural regularities in conversational speech actually support this hypothesis?

Parts-of-speech tags are often used to label the various categories that can be extracted from the abstract structure of languages. Different tag sets are used for languages which differ in structure, and the extent to which tags capture detail varies with the particular context in which tagging is employed. These tags are assigned automatically by statistical tools, typically assuming a Markov process, which employs regularities in word co-occurrence patterns over word sequences of varying sizes (Collins, 2002; DeRose, 1988). The fact that taggers achieve high levels of accuracy suggests in turn that high levels of systematicity must be present in distributional patterns. That is, the fact that structural properties of the training set will translate to novel and larger samples implies that the captured properties are sampling invariant. Previous work on text corpora implies that, in text at least, the empirical distributions discriminated by communicative contexts are geometric (Ramscar, 2019). This raises a question: do the patterns that emerge during part-of-speech tagging also discriminate distributions with similar empirical properties?

Further, the finding that the probability of types that are subcategorized by these context decreases at a constant rate (Ramscar, 2019) suggests in turn that different empirical subcategories might serve similar communicative purposes at different levels of specificity. In English, message length in words has been shown to increase as the content of messages increases as a consequence of learning and specialization (Borensztajn et al., 2009; Gil, 2008; Klingenstein et al., 2014). The apparent systematicity revealed by analyses of covariance patterns in text suggests that communicative codes may be adapted to support the transmission of an unbounded set of messages at multiple levels of description, including length. That is, in speech at least, the considerations reviewed above would seem to suggest that

word sequence length (at least in English) may be related to the relative probability of the message with respect to all messages all speakers might want to communicate. This raises a further question: is the distribution of n-grams in speech geometric?

To answer this, we analyzed the distributions of part-of-speech labels, utterance length, and utterance position in the Buckeye Corpus of conversational English (Pitt et al., 2005).

3.3.1 Results

Figure 3.2a–c shows frequency rank distribution plots of log counts for part-of-speech labels, phrase lengths, and the phrase positions, respectively. The blue line indicates the best fit to log-log scale (power law), while the red line shows the best fit to geometric (geometric is linear; the probability decreases at a constant rate). As we can see, the empirical distribution (represented by the grey points) of part-of-speech labels, utterance length, and utterance position, with R^2 of 0.9725, 0.9957, and 0.9981, respectively, show a close fit to geometric, whereas fits to power law are 0.7798, 0.8109, and 0.8035, respectively.

These results thus suggest that sampling from the functional distributions that can be discriminated by context at this level may indeed result in probability estimates that are similar across speakers, irrespective of discourse context and length.

In addition, they provide some evidence to support the suggestion that hearing a one-word utterance such as *yes*, *okay*, *correct*, or *exactly* or a longer utterance such as *um sort of let them make their own decision when they got older what they wanted to do* is sufficiently stable irrespective of size, again indicating that the distribution of communicative types at different levels of description in conversational speech may be systematic.

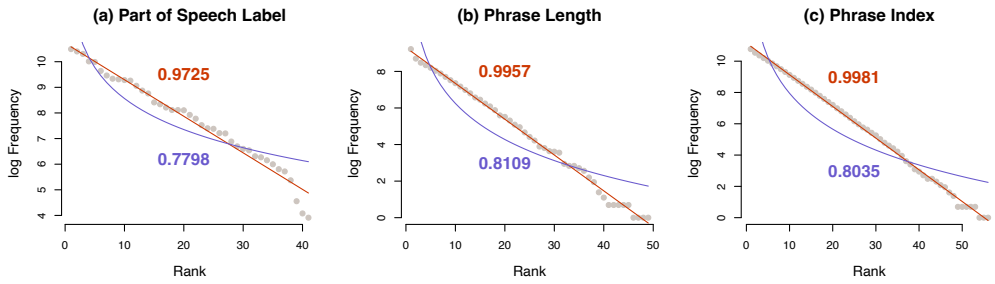

Fig. 3.2: The frequency distributions for the part of speech label (**a**), utterance length (**b**) and utterance position (**c**) categories in the Buckeye Corpus (Pitt et al., 2005): Grey points show the observed distribution, with fits to a power law distribution (blue line) and a geometric distribution (red line). All three distributions show a close fit to a geometric distribution.

3.3.2 Discussion

Our results show that grammatical subcategories captured by part-of-speech tags have distributions that are likely to lead to an alignment in the probabilistic expectations of speakers regardless of any differences in their exposure to these distributions. They also provide further support for the suggestion that, unlike aggregate word token distributions (which have power law distributions (Baayen, 2001)), the empirical distributions that are discriminated by communicative context are geometric (Ramscar, 2019).

The abstract model of communication defined by Shannon (1948) is at heart a deductive process of uncertainty reduction (Ramscar, 2019). The model assumes that communicative codes will be distributed so as to ensure that every sequence produced has the same statistical properties. A consequence of this is that any mixture of code samples will have the same statistical properties as any other sample. By contrast, it would appear that, in speech at least, natural languages gradually reduce message uncertainty via a series of sequential subcategorization frames of increasing degree of specificity. Evidence for this suggestion is provided by the differences in type/token ratio of part-of-speech categories, which vary systematically. Further, the shape of the distribution of utterance lengths suggests that the expectations about the distribution of messages of different lengths that speakers learn will likely align, helping the overall system to deal flexibly with the ever-growing number of specific messages that humans are likely to wish to communicate.

In the introduction, we described how constraints on the structure of name sequences have lead to qualitatively different patterns of distribution in English and Sinosphere first names. Legal constraints on last names in English have lead to differentiation between (geometric) local first name distributions which, when aggregated over, fit power laws (Ramscar, 2019). Thus, the differences in the extent to which word categories are subjected to grammatical and lexical constraints (Section 3.1.1) seem to predict differences in the productivity of lexical categories over time, leading to more aggregation in verbs. The analysis presented in the next section aims to explore whether the shape of word frequency distributions of different lexical categories reflect the differences in the way they are constrained by the grammar.

3.4 Word Distributions across Lexical Categories

As we noted above, high-level descriptions (e.g., parts-of-speech) clearly capture many abstract communicative properties such as animacy, agency and number in nouns or tense, and aspect or argument structure in verbs. However, it seems

that the functionality of these categories is further subcategorized by patterns of co-occurrence which encode more specific distinctions between agents, objects, actions, and relationships. This implies that verb and noun frequency distributions are aggregates of functionally distinct subcategories. Consistent with this, Bentz et al. (2014) show that aggregates over verbs and nouns are power law distributed while Ramscar (2019) confirmed this finding and then showed that the subcategorical distributions of verb and nouns discriminated by communicative context are geometric.

Importantly, previous studies have shown that token distributions in closed class categories (function words and modal verbs) do not follow power laws (Bentz et al., 2014; Piantadosi, 2014). These departures from the trend to power law in other categories are assumed to be related to the communicative function of high-frequency words. Linguistic theories typically assume that closed class tokens serve a qualitatively different modifying or grammatical function while open classes are considered to contain and transport meanings; that is, they provide lexical contrast.

These previous results thus predict that, when context is not used to subcategorize them, nouns and verbs in English will be distributed differently to function words. To explore these patterns of distribution, we analyzed the word token distributions of these separate parts of speech across the speech samples.

3.4.1 Results

There are 44,722 noun and 45,159 verb tokens in the analyzed sample. With 5817 unique types, nouns are a far more lexically diverse category than verbs with 2574 types. By contrast, the 116,960 function word tokens are represented by 144 unique types.

Figure 3.3 shows the token distribution of the three largest grammatical categories. We can see that both verbs and nouns have a closer fit to power law compared to geometric distribution: $R^2_{pl} = 0.976$ and $R^2_{geom} = 0.701$ for verbs; $R^2_{pl} = 0.971$ and $R^2_{geom} = 0.772$ for nouns.

By contrast, the 144 unique function words ($n_{tokens} = 11,696$) show an almost perfect fit to geometric $R^2_{pl} = 0.796$, $R^2_{geom} = 0.992$. A separate analysis shows a better fit to geometric over power law in distributions of determiners ($n = 16$, $R^2_{geom} = 0.953$, $R^2_{pl} = 0.830$), pronouns ($n = 28$, $R^2_{geom} = 0.957$, $R^2_{pl} = 0.741$), and prepositions/subordinating conjunctions ($n = 78$, $R^2_{geom} = 0.983$, $R^2_{pl} = 0.863$). The aggregated set of function words, however, improves the fit to geometric.

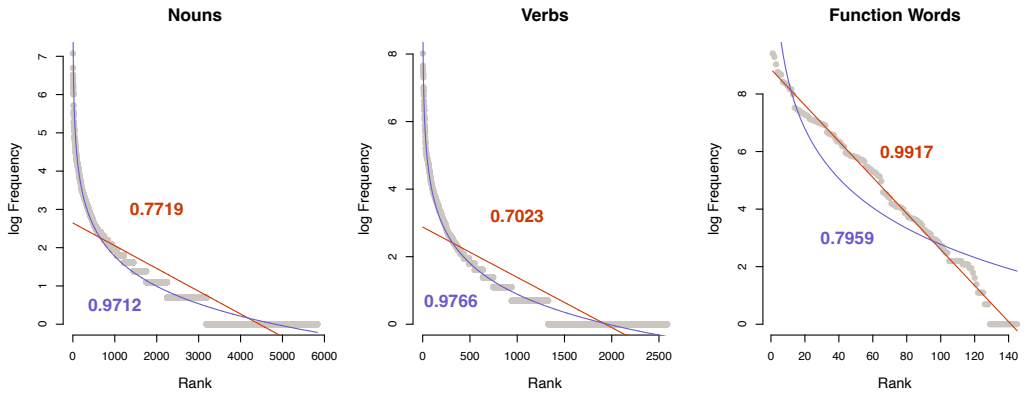

Fig. 3.3: Word frequency distributions of nouns, verbs, and function words in the Buckeye Corpus (Pitt et al., 2005) show that the substantially smaller (compared to nouns) set of verbs has a closer fit to power law distribution, indicating more aggregation. The shape of the distribution in function words suggests that function words form a natural empirical distribution.

3.4.2 Discussion

When taken in conjunction with earlier findings (Bentz et al., 2014; Piantadosi, 2014; Ramscar, 2019), the distribution of function words we observed here supports the suggestion that they form a natural communicative distribution. This in turn suggests that, despite the fact that prepositions (in contrast to determiners and pronouns) distinguish between spatial and temporal relations, prepositions, determiners, and pronouns are part of the same functional subsystem and, at some level, serve the same communicative function.

By contrast, we find that the lexically more diverse categories fit power laws. As previously discussed, these distributions could be the product of aggregating over multiple communicative distributions serving distinct communicative functions. This suggestion is further supported by the observed distributions of verbs and nouns, which suggest that a smaller number of unique verb types appears across a larger number of distinct communicative contexts than is the case for nouns. This observation is supported by the fast growing head in the verb distributions, which appears to result from aggregating over high-frequency verbs, whereas the fast growing tail in the noun distributions appears to reflect the greater volume of low-frequency nouns.

In other words, the results imply that the differences observed between lexical categories do not necessarily warrant categorial distinctions. Rather, the observable differences appear to reflect the extent to which word co-occurrence clusters are shaped by the opposing communicative pressures of prediction and discrimination over the course of learning.

In other words, these results confirm the idea that lexical categories are not equally distributed across utterance positions. The next part of our analysis explores these relationships further.

3.5 Lexical Category, Word Order, and Recurrence Patterns – What Makes a Lexical Category?

The distribution of function words suggests that function words will form the grammatical subcategory that is first discriminated systematically from the speech signal. As a consequence, it seems likely that, as both intuition and many linguistic theories would predict, function words provide a first contextual frame to aid in the learning of other grammatical and contextual categories. Once these basic contextual frames are learned, they will provide context, assisting in the learning of other words. The idea that context will provide information that aids learning in turn suggests lexical diversity will increase with utterance length.

Consistent with this suggestion, Genzel and Charniak (2002) have shown that, although caching local probability estimates of a words' occurrence in written samples (to account for the variance in recurrence patterns over time) stabilizes relative entropy over lexical sample size in nouns significantly, the effect is far smaller in verbs and absent in function words. In the light of the foregoing discussion, this might be taken to suggest that patterns of co-occurrence in verbs are less variant than those in nouns and that these patterns are still less variant (and may even be regular) in function words. These considerations suggest in turn that the different subcategories of words systematically reduce uncertainty in communication at different levels of abstraction. To explore whether the different communicative contributions of words from different lexical categories are quantifiable in the speech signal, we analyzed the patterns of occurrence of nouns, verbs, and function words (the three largest categories by token count) over utterance length.

3.5.1 Results

The probability density of token occurrence over log normalized utterance length was analyzed by category. As can be seen in the left panel of Figure 3.4, while the larger parts of tokens in all three categories follow a normal distribution across utterance position (presumably as a consequence of utterance length), there is also evidence of distinct bursts of occurrence which align with the word order typology of English. That is, less specific pronouns are more likely in utterance initial positions,

verbs are more likely in utterance medial positions, and nouns are more likely in utterance final positions.

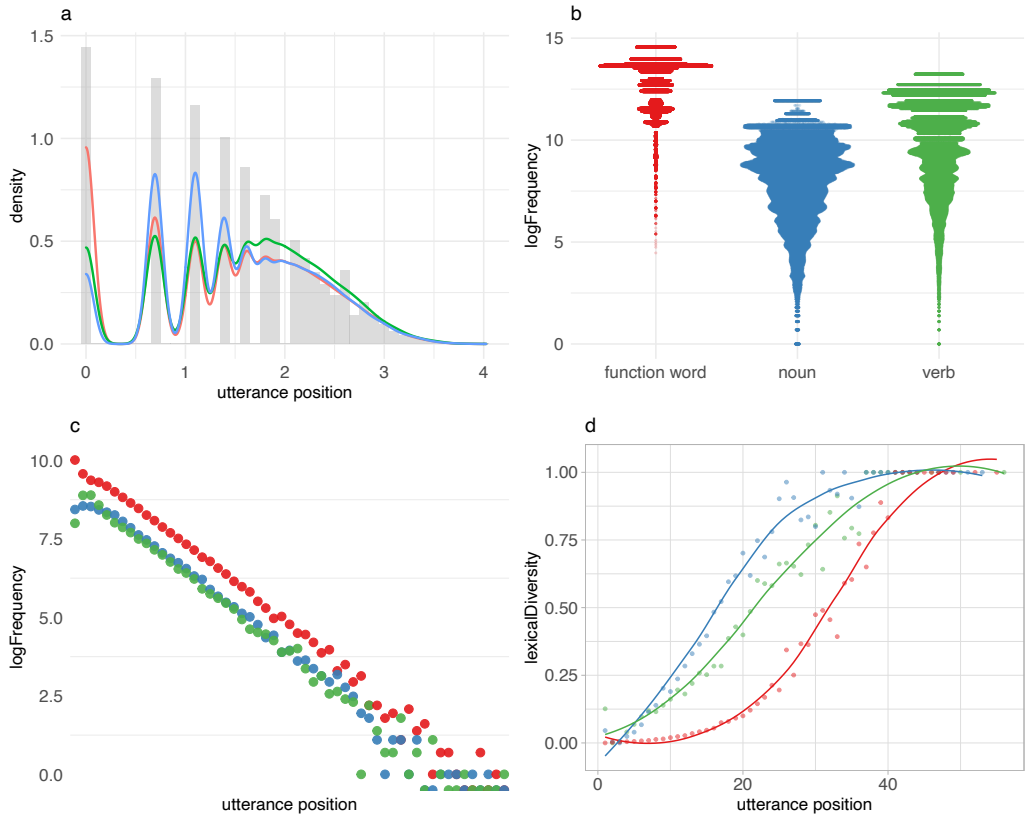

Fig. 3.4: Distributional properties of the three largest (by token count) categories analyzed by utterance position and frequency range: We find that the overall probability of occurrence varies with type and utterance position (**a**), that frequency distributions of lexical classes are not similarly distributed across the probability space (**b**), that part-of-speech token probability decreases linearly as a function of utterance position (**c**), and that lexical diversity increases nonlinearly as a function of utterance position (**d**).

The extent to which lexical categories are represented across the probability space (Figure 3.4b) is correlated to the average utterance position. We find 85% of all function word types in the top 50 tokens, which makes up 51% of the probability mass, and 93% of function word types in the top 100 words, which makes up 64.6% of the probability mass. In other words, function words are high-frequency words.

Further, we observe that, while token probability across lexical categories decreases linearly (Figure 3.4b) over utterance position, the increase in lexical diversity across all three categories is nonlinear. The right panel of Figure 3.4 shows smoothness of the normalized type/token ratio as a function of utterance position. We observe significant differences in the patterns of increase between the three lexical classes. The increase in the lexical variety of function words is limited to a small number of tokens in the latter positions of long utterances. The diversity in nouns increases earlier than in verbs.

Figure 3.5 shows that when words at utterance boundaries are excluded from the analysis, the normalized type/token ratio of nouns and verbs show similar increase patterns while the growth in function words remains unaffected. In contrast, the wide confidence interval in utterance final verbs indicates that the relationship between lexical diversity and utterance length (which can be taken to signify context) is less consistent in verbs than it is in nouns (and pronouns).

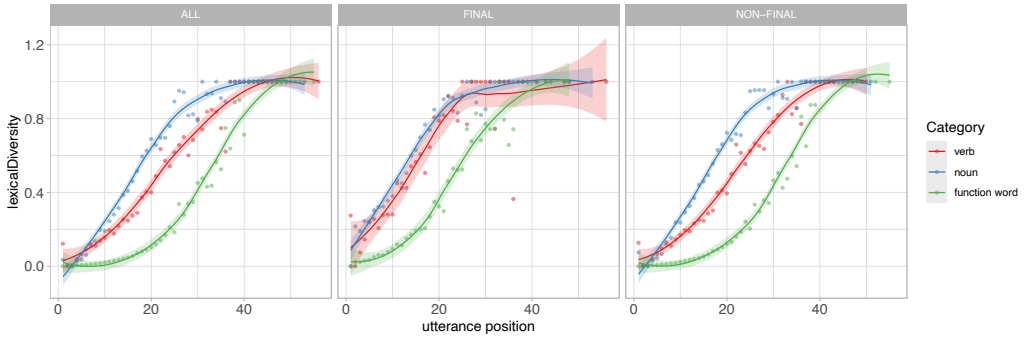

Fig. 3.5: Increase in local lexical diversity (type/token ratio) across utterance position is not linear. The increase rates differ substantially between lexical classes. The differences in the increase rate between verbs and nouns in utterance final position are restricted to utterance initial tokens. The confidence interval in verbs is larger. The differences in the increase rate between verbs and nouns are constituted by the extent to which context affects lexical variety in nonfinal tokens.

3.5.2 Discussion

In sum, we observe significant differences in distributional properties between word categories. Word categories differ with respect to the frequency range they populate, the average utterance position, and lexical diversity. From the perspective of learning, this implies that the properties which distinguish word categories interact with the order in which they are learned while the order in which they are learned appears to be a consequence of the regularity with which they are represented across samples.

We suggest that the aggregation effects in token distributions across lexical classes is correlated to the degree in which category types are regularly distributed across language samples, reflecting the extent to which their communicative function is mediated by the contextual frames they appear within. Our analysis shows that lexical classes differ both in the average utterance position and in the rate at which lexical diversity increases as a function of utterance position and that the increase rate is inversely related to the average utterance position of the class.

In the next part of the analysis, we explore the extent to which the variety of abstract grammatical constructions in which words are embedded can serve to capture the

differences in distributional structure and recurrence patterns that we observe across lexical categories.

3.6 Distribution of Grammatical Context – How Do Different Parts of Speech Carry Out Their Communicative Function?

Words often occur in multiple grammatical contexts. The word *claim*, for instance, appears 5989 times in the spoken section of the Corpus of Contemporary American English (Davies, 2010): 2719 times as a verb and 3270 times as a noun, 3016 times as noun singular, 1994 times as base form verb (1), and 1276 times as an infinitive (2), so that the three instances of *claim* in the three following examples serve distinct communicative functions which are not equally probable.

11. The girls *claim* to have seen the fairies.
12. You may be able to *claim* compensation.
13. The court found no evidence to support her *claim*.

The particular uses of *claim* that speakers intend to communicate will thus be determined by the lexical and grammatical context in which it is used. If one were to count the word *claim* as one type across all the contexts it occurs within without taking into consideration its lexical status, one would run the risk of aggregating over the multiple communicative functions it serves, and this problem will clearly increase as a word's frequency increases.

To explore the extent to which lexical subcategories receive support from these kinds of contextual frames, we next analyzed the distributional properties of the frames that words are embedded within by lexical category.

3.6.1 Results

In the first part of this analysis, we explored the distributions of grammatical context (defined as part-of-speech bigram) that words are embedded in. This was then followed up with an analysis of the word token distributions that these part-of-speech constructions distinguish between.

The left panel of Figure 3.6 presents the log frequency rank plots of part-of-speech bigram distributions over the three analyzed lexical categories. It shows that all

three distributions are geometric and that the slopes differ substantially. The slopes which reflect the extent to which words are subcategorized by grammatical context are inversely correlated to the rate at which lexical diversity increases as a function utterance position in Figure 3.5.

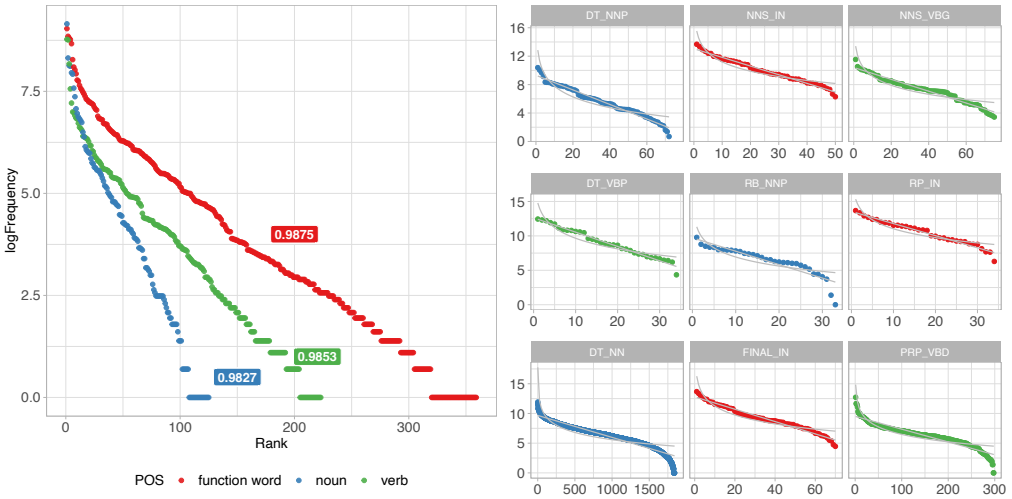

Fig. 3.6: Distribution of contextual distinctions (part of speech bigrams) by lexical class: Nouns appear in a far smaller number of contextual frames; the size of the contextual frame is on average larger. The frequency distribution of verbs within the contextual frame is exponential. In the larger set of nouns, we see effects of aggregation in the low- and high-frequency tails.

We observe a more diverse set of grammatical distinctions between categories of smaller frame size in verbs as compared to nouns. the distribution of the parts-of-speech bigram of the preceding word and the word itself is more diverse, with 124, 223, and 359 parts-of-speech constructions for 5869, 3124, and 144 unique word forms, respectively. The parts-of-speech context on average comprises 37 types of verbs (ranging from 1 to 460) and 119 types of nouns (ranging from 1 to 1862). In contrast, function word contexts on average host 2 distinct function words.

This suggests that the extent to which words are subcategorized by grammatical context is correlated with both lexical diversity and the average utterance position of a category. In consequence, the lexical distributions we find embedded in grammatical frames differ in size and structure. The center and the right panels of Figure 3.6 show the frequency distributions of the unique words found in two of the smaller subcategorization frames. The smaller (by unique type count) frames show a close fit to geometric irrespective of the word frequency range. In general, we observe more aggregation in noun frames. That is, the extent to which the subsamples extracted from grammatical subcategorization frames show the effects of aggregation appears to be independent of the frequency range of lexical contrast they distinguish between. Instead, aggregation appears correlated to the size of the subsample and,

by implication, the extent to which lexical frames serve further subcategorization within the more abstract grammatical frames.

3.6.2 Discussion

The distribution of grammatical constructions suggests that nouns which on average appear in a smaller number of more lexically diverse constructions will receive more support from lexical frames, resulting in less variance in the conditional probability between nouns and the words on which they are conditioned. That is, the more high-frequency nouns tend to appear in larger, high frequency contexts and thus tend to be further subcategorized by smaller lexical subcategorization frames. In contrast, the extent to which the variety of grammatical contexts serves to reduce uncertainty across a smaller (by type count) set of verbs will lead to more variance in the conditional probability between verbs and verb arguments.

In the next chapter, we explore the effects that the distinct patterns of covariance between high-frequency verbs and high-frequency nouns and their collocates have on the variety of articulated variants we find in the speech corpus.

4 From Information Structure to Speech Form

The results we have described in chapter 3 suggest that speech sequences are structured to allow for for efficient message transmission over multiple nested levels of description. The distribution of words and parts of speech indicates that information structure *depth* increases over message sequences, supporting gradual increases in the degree to which low-level sublexical contrasts (e.g. the duration and acoustic properties of syllables and phonemes) contribute to resolving uncertainty about messages. Consistent with this, it has been shown experimentally that speech rates are perceived as being faster and target words as being longer when cognitive load is increased (Bosker et al., 2017), a response pattern that suggests that speakers adapt their response to the relative uncertainty resulting from utterance context.

The notion that sensitivity to speech signals resolves at multiple timescales and that this variances seems to reflect adaptation to uncertainty is consistent with evidence that sublexical variation in speech sequences increases with sequence length and position. This phenomenon is characterized by the strengthening of word initial consonants and the lengthening of final vowels. While both effects increase cumulatively as a function of utterance length (Fougeron and Keating, 1997), the interaction between lengthening and strengthening is weak, indicating that hyperarticulation and vowel space expansion are not equally affected by context. Moreover, while low-probability and word initial segments are more likely to be stressed and while segment deletion is more likely in high-frequency phonemes and in latter positions, the frequency effects actually observed in very frequent segments depart from this pattern. Also, the correlation between duration and extreme articulation, and duration and frequency declines as a function of utterance position (Van Son, Pols, et al., 2003b). What does this tell us about speech form variation in context?

The analyses presented in Section 3.6 indicate that average grammatical uncertainty peaks in words that are more likely to occur in utterance initial positions and that average lexical uncertainty peaks in categories which are more likely at utterance final positions. It has also been shown that slow-downs in articulation are associated with uncertainty and that uncertainty leads to an increase in articulatory variance. These effects have been observed both within (Salverda et al., 2003) and across word

boundaries (Bell et al., 2003) as a consequence of syntactic irregularities (Tily et al., 2009) and appear functional in lexical decision (Kemps et al., 2005) and discourse (Arnold et al., 2007). Since our analyses show substantial differences across parts of speech in both the extent to which words are predicted by the previous context and the extent to which they serve to predict the upcoming part of the message across the frequency range, this seems to imply that the apparently inconsistent effects of frequency that have been previously observed are both predictable and systematic with respect to the structure of the grammatical context.

This in turn can be taken to suggest that sublexical variance follows as a consequence of an increase in lexical and grammatical variety in which words are embedded and that the variants we observe aim to increase the efficiency in transmission of informative contrast at multiple levels of description. To examine this question we conduct a statistical analysis of the effects of variation in the collocations of words on the number of distinct forms found in the speech corpus. The results are presented in the next section.

4.1 Effects of Frequency and Collocate Diversity on Variation

4.1.1 Disentangling the Effects of contextual diversity and frequency

Wedel and colleagues have shown that the number of competing minimal pairs in lexical context predict likelihood of vowel merger (Wedel et al., 2013b) and voice onset time duration (Wedel et al., 2018), suggesting that what drives speech contrast loss is the extent to which minimal pair competition is resolved in context. In line with this, Piantadosi et al. (2011) observe that the relative probability of a word in a lexical context (defined as word sequences ranging between 2–4 words) is a far better predictor of word length than word frequency.

This raises questions: Does this hold for variance too? Is the diversity of collocate contexts across which a word appears a better predictor of the extent to which a type will vary across a speech sample than frequency?

The probability of a known word appearing in a previously unattested context increases with the average word count so that word frequency and collocate diversity are strongly correlated ($r(9190) = .70$, $p < .0001$). High-frequency words are more likely to be preceded by a larger number of different words and thus tend to appear

across a larger number of communicative contexts that vary in size. By implication, there is more variance in the conditional probability between high-frequency words and their collocates. In contrast, words from the mid-frequency range will appear in a smaller number of distinct communicative contexts, leading to less variance in the conditional probability between mid-frequency words and their collocates. In line with this, an analysis by Arnon and Priva (2014) shows that, in contrast to results reported by Bell et al. (2009), duration in content words is affected by both word and multiword frequency as well as the transitional probability of both the preceding and following collocates when high- and low-frequency trigrams; sequences interrupted by pauses and word final sequences are excluded from the analysis. Finally, the increase in lexical diversity over utterance length (Section 3.5) suggests that low-frequency words tend to appear in a larger number of distinct message contexts, again leading to more variance in the conditional probabilities of low-frequency words at different positions within the sequence with respect to the likelihood of the message.

The discriminative nature of learning predicts that this variance will increase within-context competition over exposure time and that this will minimize the informativeness of contextual cues which predict a large number of lexical contrasts. This in turn predicts more sublexical variation in words that serves as cues to a larger number of collocates, reflecting the uncertainty of the relative context. Taken together, these factors predict distinct patterns of variance across frequency ranges.

4.1.2 Results

To explore the nonlinear effects of frequency and collocate diversity on observed variance, we fitted generalized additive mixed models (GAMM) (Wood, 2017) using the *mgcv* package for R. In baseline model 1, we model the normalized number of observed corpus variants as a function of the smooth over log frequency. In baseline model 2, we model the number of variants as a function of a smooth over collocate diversity, the log normalized number of preceding words we observe in the corpus. Model 1 counts show a strong, nonlinear effect of frequency ($p < 0.0001$). It yields an R^2 of 0.435 and explains 43.5% of the deviance in the data ($edf = 5.05$, $AIC = 19852.04$). Model 2 shows a strong, nonlinear effect of diversity of collocates in the preceding position ($p < 0.0001$), explaining 74.6% of the variance in the data ($edf = 6.922$, $R^2 = 0.746$, $AIC = 11777.66$).

We assessed the goodness of fit of both models by the Aikake Information Criterion (AIC). Model 2 improved the score by 8074.38. To contrast the contribution of both predictors, we modeled word variance as a function of smooth over log normalized word frequency and log normalized number of variants observed in the corpus in a

combined model 3. Model 3 ($R^2 = 0.746$, $AIC = 11548.74$) reduced the AIC by 228. Both predictors are highly significant ($p < 0.0001$).

Interestingly, the plots show that the frequency effects predicted by the baseline model 1 and the combined model diverge substantially across frequency ranges (see Figure 4.1a,c), suggesting that the effect of frequency is largely overestimated in the low-frequency and mid-frequency ranges by the baseline model. It further appears that a large part of the frequency effect is confounded by the correlation between word frequency and the number of collocate contexts a word appears within. There remains, however, a strong effect of frequency observable in high-frequency words. The high-frequency part of the data behind the effect comprises 82 function words, 57 nouns, and 47 verbs, representing 69%, 1%, and 2% of unique types, respectively.

Word frequency appears to influence the extent to which a word varies in form only in high-frequency words and thus holds for type variation across lexical categories to the degree with which the category is represented in the high-frequency tail of the word distribution. We further observe a stronger correlation between collocate diversity and word frequency in function words ($r(143) = 0.882$, $p < 0.0001$), than in verbs ($r(2549) = 0.665$, $p < 0.0001$) and nouns ($r(5626) = 0.593$, $p < 0.0001$).

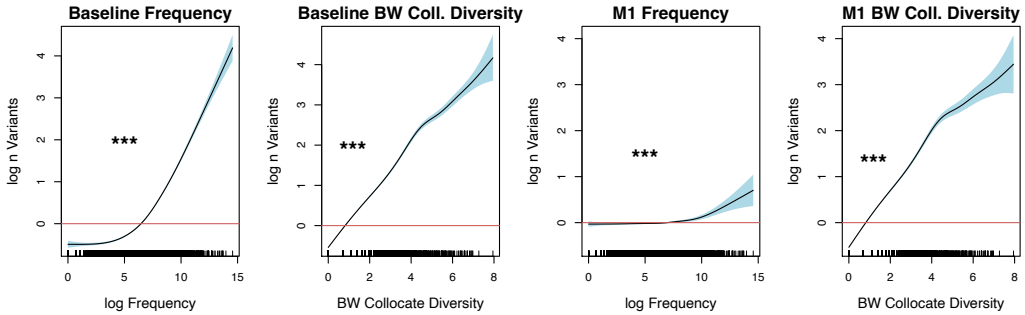

Fig. 4.1: Baseline word variance model comparison: (**a**) the log normalized number of observed variants as a a function of smooth over log frequency (derived from the spoken part of COCA); (**b**) the log normalized number of observed variants as a function of collocate diversity, the log number of preceding words; and (**c**,**d**) Figure 4.1a,b in a combined model.

Finally, we fitted a set of combined models, adding in the log number of distinct parts of speech following each word for all words (model 4) and adding in lexical category as a covariate factor (model 5). In model 4, we observe a fairly weak effect of frequency ($p < 0.006$ (see Figure 4.2a), while the effects of the context predictors were strong. The AIC score is reduced by 531.

In model 5, the introduction of lexical category as a covariate further reduces the AIC score by 254 points and explains 76.8% ($R^2 = 0.766$) of deviance observed in the data. The effect of frequency is not significant in verbs ($p < 0.816$) and function

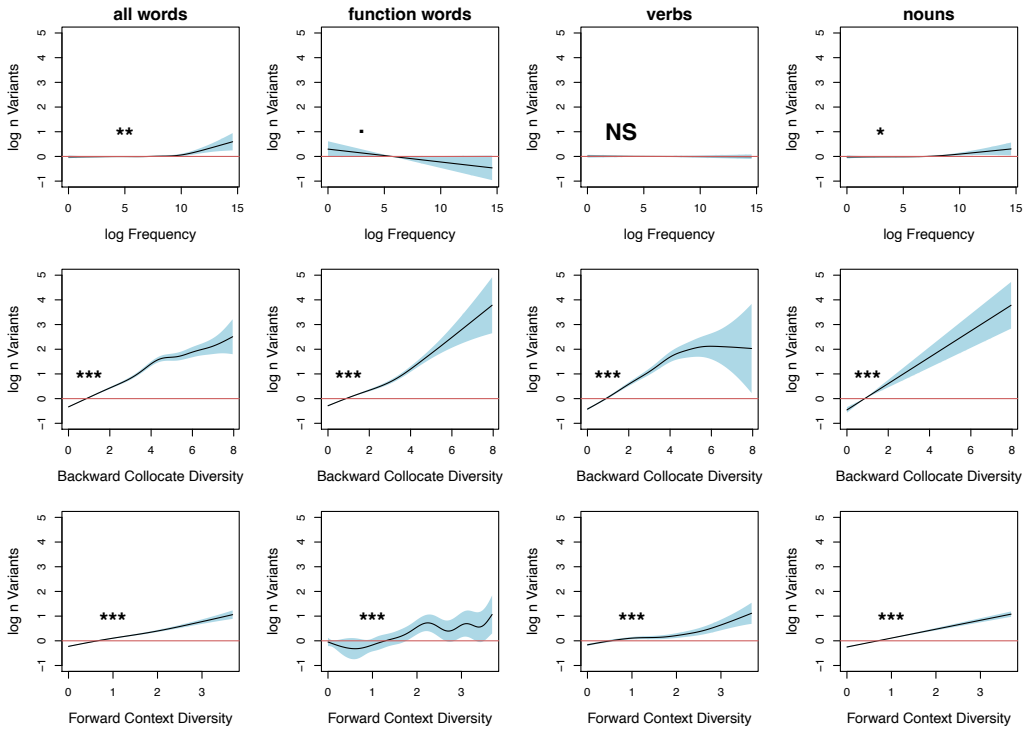

Fig. 4.2: Log normalized number of observed variants as a function of smooth over log frequency (row 1) and adjacent token diversity (rows 2 and 3) for all words (**a**), function words (**b**), verbs (**c**), and nouns (**d**): when collocate diversity is taken into account, frequency effects on variation only hold in a minimal proportion of high-frequency nouns and appear to have no effect at all on verbs and function words.

words ($p < 0.062$) and is statistically significant but weak in nouns ($p < 0.018$). Again, all contextual predictors are highly significant in function words, nouns, and verbs. The same pattern was observed for all of the other analyzed categories apart from the following exceptions: filled pauses and numbers show an effect of frequency ($p < 0.002$); contractions are unaffected by the collocate diversity ($p = 0.21$); and there is no interaction between modal verbs and the upcoming collocate context ($p = 0.1$). Modals, numbers, and contractions comprise 0.008% of the analyzed data set. We observe differences in the effect of preceding collocate diversity between verbs and nouns in that the effect and the confidence interval both increase linearly in nouns while the effect levels off in high-frequency verbs, showing an increase in variance.

A closer examination of the data reveals that the relationship between word frequency and collocate diversity differ significantly across frequency ranges for verbs and nouns. Collocate diversity is much higher in high-frequency verbs and function words than it is in high-frequency nouns. Also, there is far more variance in the effect in high-frequency nouns.

4.1.3 Discussion

The results of this analysis align with the finding that word counts outside of their communicative context contribute little when it comes to explaining variation in articulated forms. Rather, we observe that the largest part of this variance is explained by the diversity of the lexical contexts in which words appear. The remaining effects of frequency are limited to a relatively small number of high-frequency nouns and words from closed categories (numbers, contractions, and filled pauses).

These results are thus consistent with the differences we find in the distributional patterns of lexical categories in that, unlike high-frequency nouns, it would seem that high-frequency verbs are far less likely to be encountered outside of their argument frames (supporting the idea that verbs are encountered as arguments rather than lexical items per se).

Given that our results show that the variance in the observed forms is largely explained by the covariance in the collocate structure and that patterns of covariance are systematic, this finally leads us to the question of the systematicity in the sublexical variance: Is the distribution of the observed contrast geometric?

4.2 Distribution of Word Initial Contrast

4.2.1 Why Word Initial Contrast?

Previous work on sublexical variation shows that the structure of speech sound sequences is such that the probability of speech segments at segment transitions is not independent (Van Son, Pols, et al., 2003b). Gating paradigm studies have shown that the informativeness of word medial contrast is mediated by the extent to which both the preceding sentence context and word initial phonetic contrasts have minimized uncertainty about the word (Grosjean, 1980; Grosjean and Itzler, 1984). Accordingly, the entropy in sublexical contrast peaks at word initial boundaries (Van Son, Pols, et al., 2003a). This suggests that word initial speech contrasts may serve a distinct communicative function in context.

An initial analysis of word initial phonetic label distributions over both observed and citation forms in the corpus revealed poor fits to both power law and exponential distributions, suggesting that the aggregated distribution of the phonetic labels observed in our corpus may result from mixing the underlying communicative distributions. To examine this, we used parts-of-speech classes to provide a simple, objective

method for contextually disaggregating individual communicative distributions from the mixed distribution of phonetic labels in our corpus.

4.2.2 Results

The frequency distributions of word initial phone labels were analyzed by parts-of-speech category considering the observed forms as the empirical distribution and the citation form as its model counterpart. Overall, both empirical and model distributions of phone labels show a better fit to geometric than to power-law distribution (Fig. 4.3). However, while the fits to geometric in the model distribution show a larger departure from linearity and large differences in slope between different parts of speech, the observed phones across categories converge on nearly identical distributions with close fits to geometric (Tab. 4.1).

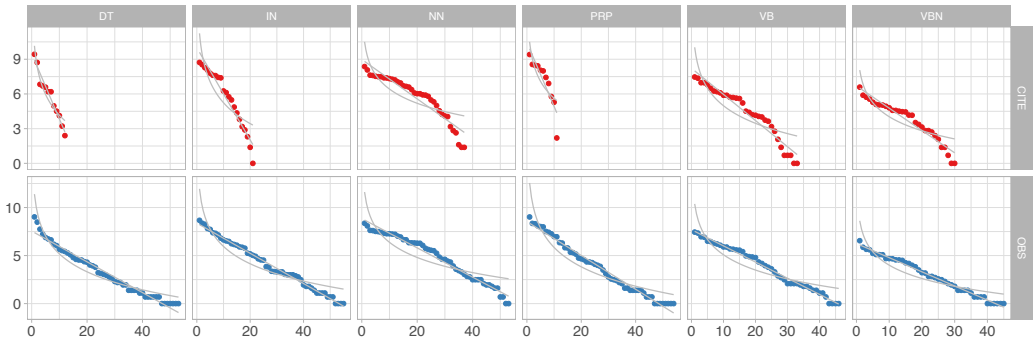

Fig. 4.3: The distribution of word initial phonetic labels in 6 selected parts-of-speech categories: Row 1 shows the distribution presupposed by the dictionary forms, and row 2 shows the distribution of phonetic variants which were actually observed.

Tab. 4.1: The distribution of word-initial phonetic labels by part-of-speech category (Penn Treebank classification) from the Buckeye Corpus of conversational speech (Pitt et al., 2005): The first two columns contain slopes from the log frequency-rank model for observed and theoretical distributions, followed by the linear model fit to log frequency - rank (R^2, geometric), model fit to log frequency - log-rank (R^2, power law) and the total number of assigned phonetic labels (n_{phon}). The model distribution represents the distribution of labels presupposed by the dictionary forms, while the empirical distribution shows phonetic contrast produced by the speakers.

Part of Speech	**Slope**		R^2_{geom}		R^2_{pl}		$\mathbf{n}_{phon}$	
	emp	$model$	emp	$model$	emp	$model$	emp	$model$
determiner	−0.16	−0.565	0.969	0.953	0.922	0.891	53	12
preposition	−0.157	−0.399	0.993	0.941	0.85	0.697	55	21
personal pronoun	−0.176	−0.554	0.984	0.802	0.87	0.593	54	11
noun, sg. or mass	−0.152	−0.171	0.975	0.895	0.726	0.621	53	37
proper noun, sg.	−0.14	−0.145	0.931	0.889	0.699	0.68	39	31
noun, plural	−0.167	−0.175	0.961	0.873	0.72	0.636	44	35
verb, base form	−0.168	−0.226	0.989	0.941	0.798	0.698	46	33
verb, past tense	−0.172	−0.244	0.972	0.944	0.784	0.768	44	31
verb, gerund/pres.part.	−0.168	−0.21	0.969	0.96	0.812	0.735	46	32
verb, past part.	−0.147	−0.193	0.991	0.934	0.828	0.716	45	30
verb, non-3rd pers.sg.pres.	−0.155	−0.249	0.991	0.977	0.777	0.751	51	31
verb, 3rd pers.sg.pres.	−0.157	−0.218	0.975	0.976	0.866	0.847	42	31
proper noun, pl.	−0.269	−0.342	0.847	0.908	0.938	0.958	12	10
function word	−0.15	−0.229	0.9731	0.8604	0.7317	0.6276	61	26
noun	−0.159	−0.199	0.9689	0.8713	0.7112	0.5797	54	40
verb	−0.164	−0.225	0.9856	0.9234	0.7406	0.6473	55	33

In both function and content words, the empirical distributions significantly improve the fit to a geometric. Importantly, despite substantial differences in the type/token ratio of the lexical classes analyzed, all of the categories have nearly identical empirical distributions with minimal differences in slopes. The exception is plural proper nouns where the data is extremely sparse (this category comprises a mere 50 tokens). Further, while we find that initial phones from several small categories (particles, modals, and filled pauses) have poor fits to either a geometric or a power law, in a similar vein, it is debatable whether these small sets of items constitute separate categories in terms of the covariate structures they populate.

Finally, we extracted time bins of initial phone duration centered by phone category to simulate an artificial set of discrete contrasts such that the simulation assumes a low-level subcategorization of phonetic contrast by duration. Again, across the parts-of-speech categories, the cumulative probability distributions of time bins show close fits to the geometric ($R^2 > 0.9662$) and poor fits to power law ($R^2 < 0.8333$).

4.2.3 Discussion

Our analyses show that word-initial phonetic annotation labels across different parts-of-speech categories approach geometric distributions. Word and segment durations approach the exponential distribution. The exponential and the geometric distribution are the two only memoryless distributions, the exponential being the continuous counterpart of the geometric, which is a discrete probability distribution. Memorylessness is the formal property of distributions where the probability of observing an event is independent of the history of the process. If duration and form distributions in context approximate the memoryless, the probability of hearing a particular phonetic-acoustic contrast or a word duration in its context does not change over time.

A memoryless distribution implies that local changes in event probabilities are not informative after exposure to a sufficiently large sample. Because the mean and the variance are equal in these distributions, distributions of speakers learning from different samples by minimizing the difference between their experience (the durations observed before) and the observation (the duration observed at a particular moment in time) will eventually converge on the mean value. These results thus suggest that what at first appears to be random variance in speech sound production may reflect a highly systematic distribution of sublexical contrasts that allows speakers to maintain sufficiently similar models or articulation.

While word initial variation is observable in all part-of-speech categories, we see that the degree to which types vary seems to correlate with the uncertainty of

the underlying category structure. Word initial articulations in verbs and nouns which receive more support from the context are not much more diverse than those presupposed by dictionary models. Articulated variation seems to lead to a redistribution of phonetic-acoustic contrast. Function word boundaries, by contrast, deviate much more. Remarkably, despite considerable differences in the degree to which initial tokens deviate from the citation form, the probability distributions of types arising from this variation converge on nearly identical distributions across parts of speech. Such uniformity can suggest that the shape these distributions take is subject to similar processing constraints.

Finally, we observe that distributions of word-initial phones assumed by the dictionary models show poor fits to both geometric distributions and power laws, indicating that, unlike words, aggregates of words from closed classes do not result in power laws. Instead, the mixtures we analyzed diverge in the mid-frequency range, approximating the Yule distribution, which appears to reflect local fluctuations in the relative probabilities of types.

4.3 Discussion

We analyzed distributions of the grammatical, lexical, and sublexical forms in conversational speech produced by 40 speakers of American English (Pitt et al., 2005) to assess the effects of the statistical structure of speech on the sublexical variance observed in the signal. Our analyses show that distributions of regularities in co-occurrence patterns, the words they discriminate between, and the articulated phonemes result in memoryless distributions. These results are consistent with previous, similar analyses of written English that satisfy many of the communicative constraints described by information theory (Ramscar, 2019). Accordingly, these results also provide further evidence that power law distributions seen in aggregate word frequency distributions are products of mixing functionally relevant/distinct distributions that are in themselves geometric (Newman, 2005; Ramscar, 2019).

The distributions in the analyzed sample suggest that, unlike the codes in artificial communication systems, human speech is a highly structured system of nested communicative distributions shaped by learning. In line with the predictions of learning theory, this suggests that speech form variation at positions of high uncertainty reflects interactions between regular structures at multiple levels of description. The variation in speech signals seems to maintain communicative efficiency by systematically increasing the contrast in signals. The empirical distributions of phonetic contrasts indicate that the variance in the pronounced forms systematically structures the uncertainty of communicative contexts. The evidence that the contexts

are structured supports the suggestion that phoneme distributions are components of a larger, highly structured communication system.

5

Finding Structure in Silence: The Role of Pauses in Aligning Speaker Expectations

> *The lack of definiteness which, from the point of view of empirical importance adheres to the notion of time in classical mechanics, was veiled by the axiomatic representations of space and time as things given independent of the senses. Such use of notions – independent of their empirical basis to which they owe their existence – does not necessarily damage science. One may, however, easily be led into the error of believing that these notions, whose origin is forgotten, are necessary and unalterable accompaniments of our thinking, and this error may constitute a serious danger to the progress of science.*
>
> — **Albert Einstein**
> Out of My Later Years

The intelligibility of speech relies on the ability of interlocutors to dynamically align their expectations about the rates at which informative changes in signals occur. Exactly how this is achieved remains an open question. We propose that speaker alignment is supported by the statistical structure of spoken signals and show how pauses offer a time-invariant template for structuring speech sequences. Consistent with this, we show that pause distributions in conversational English and Korean provide a memoryless information source. We describe how this can facilitate both the initial structuring and maintenance of predictability in spoken signals over time, and show how the properties of this signal change predictably with speaker experience. These results indicate that pauses provide a structuring signal that interacts with the morphological and rhythmical structure of languages, allowing speakers at all stages of lifespan development to distinguish signal from noise and maintain mutual predictability in time.

5.1 The problem of alignment in speech

Understanding how speakers align their expectations about the occurrence of acoustic events in speech signals is central to explaining the human capacity for vocal communication. However, the challenges this involves are easily obscured by intuitions informed by reading and writing. In contrast to reading and writing, which are self-paced and explicitly taught, speech is highly dependent on timing and is usually learned implicitly through exposure to continuous signals. Meanwhile, although the orderly and discrete way in which letters, words, and phrases appear in texts can misleadingly imply that speech is based on similar, corresponding alphabets of spoken gestures and inventories of discrete forms, spontaneous speech signals are produced by dynamic kinematic processes that guarantee a considerable amount of noise and deviation in the way speech sounds and sequences are produced, such that many of the 'acoustic segments' listeners 'extract' from speech signals are not actually present in the physical stimulus (Port and Leary, 2005; Ernestus et al., 2002). The differences between the measurable attributes of physical speech signals and their appearance to receivers pose a deep puzzle: while it is common, and at some level perhaps necessary, to talk about articulation in speech production (Goldman-Eisler, 1961; Miller et al., 1984) and segmentation in speech comprehension (Liberman et al., 1967; Cutler and Clifton, 1999), the array of findings showing that that much of what listeners 'segment' was never 'articulated' in the first place, while much of what is actually articulated is never segmented (Warren, 1970; Samuel, 2020), raise a question that has yet to be resolved: how do speakers bring order to — and make sense of — the apparent chaos?

In what follows, we approach this question as a probabilistic puzzle rooted in information theory, seeking to resolve an apparent contradiction between the formal definition of information, and the specific kind of discrete, periodic structure it requires, on one hand (see Shannon, 1948, p. 17ff), and the apparent absence of this structure in human communicative signals on the other (see Port and Leary, 2005; Ramscar, 2019; Linke and Ramscar, 2020b, for discussion). Applying the notion of information to speech is consistent with current models that treat spoken communication as a probabilistic process that relies on structured regularities in speech signals (e.g. Bell et al., 2003; Aylett and Turk, 2004; Bell et al., 2009; Tily et al., 2009; Wedel et al., 2013b; Seyfarth, 2014; Seyfarth et al., 2016; Wedel et al., 2018; Hall et al., 2018; Priva and Jaeger, 2018), such that speech perception is viewed as a probabilistic process in which hearers attempt to infer the most likely intended message from a noisy acoustic signal (Clayards et al., 2008; Mitterer and McQueen, 2009; Kleinschmidt and Jaeger, 2015).

However, this probabilistic conception of communication serves to highlight a critical difference between speech codes (and natural languages) and other information-theoretic codes, namely that the former have to be learned. This is a problem because of the highly skewed nature of lexical distributions (which, when aggregated, approach power laws) (Estoup, 1916; Zipf, 1949): the long-tailed nature of these distributions results in low-frequency words being irregularly distributed across samples, and guarantee that any individual speaker will only ever experience an incomplete sample of code (Ramscar et al., 2014; Ramscar, 2019). This, in turn, guarantees that each speaker's individual experience of codes – and hence their internal models of them – will inevitably be unique, such that any individual's experience will necessarily differ from any group average.

The variability in individual speakers' models of expectations raises a problem for the whole idea of statistical regularities in speech signals. Formally, what makes a signal a signal in information theory is that it comprises a set of structured probabilistic events that are defined by a shared code. Anything else is noise. If speech signals are signals in the same sense, then this raises the question of how speakers ever manage to learn the shared set of probabilities that structures them. In other words, how do speakers ever manage to learn to align their individual expectations about the structure of spoken signals, since this would appear to be necessary to make statistical regularities in the temporal and acoustic properties of the signal predictable and informative in the first place?

5.1.1 How do language users organize their expectations?

Part of the answer to this question lies in findings showing that speech signals that are easily understood in context often become unintelligible when they are presented in isolation (Pollack and Pickett, 1963; Bard and Anderson, 1983; Ernestus et al., 2002), which suggests that speech codes make considerable use of context. However, these findings also raise another question: what, exactly, is context? One recent, successful approach to answering this question has been to define contexts in terms of word collocations in text, using this information to operationalize the 'semantic similarities' between words as similarities between the collocational contexts in which they occur (Harris, 1954; Lund and Burgess, 1996; Landauer and Dumais, 1997; Sahlgren, 2008; McDonald and Ramscar, 2001; Gleitman, 2002). Similarly, studies of the empirical structure of word frequency distributions further support the idea that communicative context, in a broader sense, is set by systematic patterns of co-variation between informative events, such as for example speech phrases, words, syllables, and sounds. Analyses of linguistic distributions at multiple levels of description (speech segments, words, and syntactic phrases) show that learnable distributional regularities organize speech sequences in clusters of informative

contrasts that provide sites of predictable variation, the information (Ramscar, 2019; Ramscar, 2021b; Linke and Ramscar, 2020b).

These results suggest that lexical regularities are a source of information for increasing the predictability of signal sequences, indicating that lexical and grammatical forms might support the alignment process (Blevins et al., 2016; Ramscar et al., 2018). However, this suggestion simply serves to underline the problem outlined above: not only does the production and recognition of words rely on context, but in fact, it is often the case that many of the 'acoustic segments' presupposed by the acoustic model (dictionary form, that models a word produced in isolation) are not actually present in the speech signal, purplebut rather can only be inferred in context (Port and Leary, 2005; Ernestus et al., 2002). Moreover, this 'context' is not merely defined by the presence or absence of articulated parts of the signal, but also by the timing of their articulation (Dilley and Pitt, 2010; Morrill et al., 2014; Baese-Berk et al., 2019; Lamekina and Meyer, 2022). Given that the identification of individual aspects of speech signals such as 'lexical regularities' relies on the alignment of speaker expectations, it follows that these regularities themselves can only be noticed when learners have managed to extract some structure from signals in the first place. This then raises further questions: is there a consistent, time-invariant source of information to support the alignment process? That is, if learners are to be able to initially structure their expectations about speech codes to extract/infer information from signals, and if these systematic patterns of expectation are to be maintained across speakers regardless of their individual experience, it follows that some objective source of information must be available in order to facilitate this. What is it?

One possible source of information that would clearly seem to sidestep these problems is silence, or more specifically, pauses: the regular manifestations of silence in the speech signal. It has long been suggested that coordination between speakers in turn-taking is achieved by a process known as production rate entrainment (which involves the alignment of the rates at which auditory events are transmitted and processed). Variations in pause duration have been argued to play a crucial role in this process (Wilson and Wilson, 2005). Further, at least at the level of conversational turns, silent intervals across languages converge on remarkably constant averages (Stivers et al., 2009; Weilhammer and Rabold, 2003), which indicates that any variability in silent interval duration is relatively invariant in time.

In what follows, we examine whether these apparent structural regularities extend beyond turn-taking. Is temporal variation in speech pauses systematic, and can it play a role in aligning the expectations of speakers and listeners at the level of speech production and recognition as well?

Several empirical characteristics of speech pauses indicate that they could play an important role in the alignment process. First, silence – or more specifically the absence of articulated signal – is something that is clearly and 'objectively' present in the speech signal. Second, sensitivity to silence does not seem to rely on experience. Instead, this perceptual contrast appears to emerge prenatally and seems to structure other, more primitive forms of vocalization (Mampe et al., 2009; Wermke et al., 2021) and serve as a first cue to speech segmentation (Männel and Friederici, 2009; Seidl and Cristià, 2008; Holzgrefe-Lang et al., 2018). This is in contrast to sensitivity to other more complex prosodic patterns and speech sounds, which vary across languages and require more time and experience to learn. Temporal variation in the articulated and the silent parts of the signal also appears to follow distinct patterns across the lifespan. Experience leads to more individual variability in articulation rates, whereas individual variabilities in pause production decrease. In particular, as age and experience increase, the average temporal resolution of the vocal signal (speech rate) increases and becomes more variable between individual speakers and speech contexts (Quené, 2005; Jacewicz et al., 2010; Hazan and Pettinato, 2014; Tucker et al., 2021). By contrast, the duration and variability of speech pauses in conversation is surprisingly stable across the lifespan (Redford, 2013; Neuberger, 2013; Bóna, 2011; Bona, 2014; Hazan and Pettinato, 2014).

The absence of effects of experience on pause durations might seem surprising, however, these findings begin to make sense when considered against the backdrop of the mechanisms involved in their processing. In particular, the differential contribution of domain-general and domain-specific mechanisms involved in speech perception and speech production suggests that speech production and speech perception will develop differently over time (and this actually seems to be the case, see Campbell et al., 2016; Campbell and Tyler, 2018). In the brain, the discrimination of "general acoustic events" (including non-events) appears to rely on specialized auditory mechanisms which are subject to lifelong adaptation (Yan, 2003), whereas the sequential coordination of the motor processes involved in articulation (and thus articulated durations) involves more general timing circuits that subserve a variety of other capacities that involve orientation in space and time (Marien et al., 2001; Ackermann et al., 2007). Depending on prior experience, some of these latter processes appear to be 'fixed' such that they function independently of the requirements of the specific cognitive task (Krakauer et al., 2006; Wong et al., 2017). Simultaneously, it seems that in both speech perception and production, articulatory events and the variances in their execution are highly context-specific, such that the uncertainties of the contexts and the variances in the way articulations are performed within them constitute a critical part of speaker expectation (Tremblay et al., 2008; Maslowski et al., 2019). In other words, both the execution of these highly automatized motor behaviors and their perception tend not to generalize across modalities/contexts. It

follows as a consequence of these considerations that the constraints on timing that articulation imposes must in turn place boundaries on the variability of the rates at which the informative changes in speech can be produced. Accordingly, given the systematic limitations on people's ability to reliably judge noticeable differences with respect to segment duration (Friberg and Sundberg, 1995; Quené, 2007), it seems that when it comes to perception, the way in which the speech signal is interpreted across hearers must be similarly bounded. All of the above points to a likely source of misalignment in speaker/hearers: the diverging experience-driven changes in the temporal distribution of informative auditory events within the limits imposed by the converging temporal resolution of articulations.

Speech directed at children and preverbal infants is more informative in its temporal structure and contains less information in its spectral variation than speech directed at adults (Fernald, 1989; Bard and Anderson, 1983). Notably, the extent to which distinct dimensions of prosodic variation (i.e., pauses, pitch, and phrase final lengthening) inform infants' perception and segmentation appears to vary with infants' age and native language (Männel and Friederici, 2009; Männel and Friederici, 2011; Sundara and Scutellaro, 2011; Männel et al., 2013; Skoruppa et al., 2013; Sundara et al., 2015). With experience infants' attention gradually shifts away from prosody toward more complex structural regularities in speech signals. However, pauses seem to remain a critical cue to structure throughout childhood and across the adult lifespan (Mueller et al., 2008; Peña et al., 2002; Männel and Friederici, 2016). Whereas pause durations and the syntactic structures in which they occur do not seem to change substantially across adulthood, the way older speakers realize words does appear to become increasingly context- and speaker-specific (Lieberman et al., 1989). Critically, older speakers' vowel realization, both in terms of duration and the relationship between duration and formant dispersion, becomes more variable in connected speech (Munson et al., 2011; Fletcher et al., 2015; Gahl and Baayen, 2019). Notably, these effects do not transfer to experiments where words are produced in isolation (see Watson and Munson, 2007). This is important because vowels account for the largest part of the variability in word and syllable durations in English, and these developments seem to suggest that the temporal resolution of conversational speech signals at the word and syllable levels change continuously across the lifespan. Conversely, higher-level regularities such as syntactic phrasing and phrase-level segmentation remain relatively stable.

This general idea – that learning inevitably changes individual models of the world – is further supported by studies of memory performance in healthy older speakers (Ramscar et al., 2013d; Ramscar et al., 2017). The processes revealed in these studies seem to guarantee that aspects of the signal that are informative in early communicative experience (e.g., timing, prosody, phrase) will become increasingly uninformative (i.e., stable) as experience grows, and sensitivity to the information

provided by fine-grained articulatory variation increases. These differences lie at the heart of the problem of alignment in speech. The idea that speech signals are informative is largely uncontroversial. However, formal measures of information ultimately rely on the existence of objective 'events', predictable changes in the temporal and acoustic properties of the signal that are used consistently by all speakers. As we have outlined above, the rates at which information is extracted from the signal will change with speaker experience. Taken together, the considerations above point to an obvious problem when it comes to speech information: because what counts as an informative speech event will vary with experience, most articulated parts of speech signals do not seem compatible with any formal notion of information.

These problems appear to stand in opposition to the idea that the information provided by phonemes, morphemes, or words can play a central role in speaker alignment. Instead, they suggest that speakers must first achieve alignment in relation to some objective rate at which events occur in speech, in order for them to adapt their expectations and hence be able to extract the informative events (and the systematic relations between them) from the signal (Edeline and Weinberger, 1993; Morrill et al., 2014; Finn and Hudson Kam, 2015; Xie et al., 2017). Given that articulations vary both physically and temporally, and that human vocal communication (and learning) rely on speakers' ability to anticipate these events in time (Patel, 2006; Patel, 2021), it seems that an aspect of the signal that is stable in at least one of these dimensions is required for these processes to occur. Which brings us back to pauses: can they provide this stable source of information in the speech signal?

As we noted above, a basic prerequisite of an alignment signal is that the information it provides must be invariant across speakers and available to learners of all levels of experience. Formally, in a community of speakers whose shared experience of a class of events differs, alignment will only be possible if the rate at which the events from this class reoccur is somehow independent of the degree to which speakers have sampled those events (i.e., if the results of sampling are somehow independent of experience). Yet all and any of the articulated parts of the speech signal appear to be anything but invariant. Accordingly, it follows that if pauses, which by definition are not articulated, actually do serve as an alignment signal, then we should expect them to be distributed in signals in such a way that the information communicated by them will be independent of speaker experience (within some minimum bound of experience (Shannon, 1948)). Critically, given that the properties of distributions are determined by both how we define their constituents (what is measured) and their boundaries (how the observation space is limited) it is important that we be clear about what we mean by 'speech pauses' before we proceed with our analysis.

5.2 What are pauses and what do they do?

Not every silence associated with speaking is a pause. Moreover, it is clear that even the events we might theoretically term "pauses" do not form a single coherent class. Previous work has shown that the durations of silent intervals in speech follow a log-normal distribution, which in turn appears to be the product of aggregating (at least) two distinct components: **gaps**, silent intervals between speaker turns, and **pauses**, silent intervals within speaker's turns (cf. Heldner and Edlund, 2010). With regards **pauses**, studies of speech corpora show that pause distributions are bimodal or trimodal (Demol et al., 2006), with analyses suggesting that they cluster at around 150 ms, 500 ms, and 1500 ms (Campione and Véronis, 2002), providing empirical support for the traditional classification of pauses into short (< 200 ms), medium (< 1000 ms) and long (up to 3000 ms in spontaneous speech). However, while the nature of the processes that give rise to these classes is poorly understood, what is important for current purposes is that it is possible to make a functional case for this classification.

There is evidence that pauses affect the ease with which speech utterances are processed. Importantly, pauses of different duration affect sentence processing differently: while pauses below a threshold of around 200 ms are not explicitly detectable (Walker and Trimboli, 1982), these very short pauses do appear to improve speech recognition. By contrast, the absence of silent intervals or unusually long pauses make sentences harder to comprehend (Fors, 2015). More generally, however, while it is clear that they provide crucial information about the relations between events in speech processing, pauses have also been subject of considerable study in relation to learning and error processing in the brain. As highlighted above, feedback adaptation – an important aspect of articulation – involves multiple neural systems whose individual contributions are modulated by the durations of intervals between events (Teki et al., 2011; Diedrichsen et al., 2005; Lewis and Miall, 2003; Buhusi and Meck, 2005; Coull et al., 2011), which in turn affects the quality of the learning outcomes (Foerde and Shohamy, 2011; Baese-Berk and Samuel, 2022).

In auditory learning experiments, which like speech perception, involve dynamic reorganization of temporal (when) and auditory (what) expectations through exposure to structured sequences, the specific ranges of interval durations also appear to influence what gets learned. Different delay durations appear to serve as cues to discrimination over the two different sources of information in tone sequences: tone frequency and tone duration. Talking clearly involves more than discriminating between tones of different durations, and speakers seem to be less sensitive to changes in interval duration of speech sequences than tone sequences (and presumably more sensitive to changes in spectral quantities) (Grondin et al., 2011).

However, the mechanisms involved in temporal discrimination of both speech and tone sequences appear to be surprisingly similar, and adaptation to the temporal dynamics of auditory sequences appears critical to speech segmentation and, by consequence, speech intelligibility. Given these parallels, it is notable for current purposes that results from auditory discrimination tasks indicate that inter-event intervals (pauses) shorter than 250 ms and longer than 750 ms only inform temporal discrimination. By contrast, intervals between 250 and 750 ms aid in the discrimination of frequencies as well (Buonomano et al., 2009). This qualitative divide in the information provided by shorter and longer pauses is also evident in the differences in people's sensitivity to small time perturbations in intervals (both within and between tone sequences). In particular, *just noticeable differences* in durations of intervals shorter than 240 ms are independent of the actual durations themselves, whereas*just noticeable differences* in intervals longer than 240 ms are a function of the durations to be distinguished (Friberg and Sundberg, 1995; Repp and Su, 2013).

In relation to speech, the results summarized above indicate that differences in the way pauses from different duration ranges are experienced may affect the way subsequent changes in the signal are detected. Consistent with this, it has been shown that the articulation rate of the earlier parts of utterances co-determines which words and speech segments listeners extract from the speech signal, such that shifting these rates can change what listeners actually hear/extract (Dilley and Pitt, 2010; Morrill et al., 2014; Baese-Berk et al., 2019). These effects of speech tempo on speech perception are likely a function of the way variance in the signal influences the way listeners experience the durations of consecutive sounds in sequences. Studies have shown that subjective tone durations stretch or shrink in relation to the durations of the preceding tones in a sequence (Nakajima et al., 1992). Durations of subsequent tones are overestimated if the tones immediately preceding them are considerably longer, whereas tones that follow markedly shorter precedents are experienced as being shorter, with a consequence of this effect being that speakers can experience consecutive sound intervals whose durations are objectively different as being the same (Hoopen et al., 2006). However, the introduction of silent intervals between tones leads to a weakening or an inversion of this effect (Sasaki et al., 2010), a finding that also transfers from the isochronous sequences used in experimental contexts to speech, where pause insertions have been shown to increase the intelligibility of time-compressed speech (Ghitza, 2011).

Our hypothesis is that alignment is achieved through the establishment of shared expectations about the rate at which informative events will occur, because, by definition, this is a fundamental requirement for speech events to be "informative". Accordingly, we suggest that the intelligibility of speech after linear time-compression (Ghitza, 2011) results from a misalignment between the rate at which informative

events are expected and the rate at which events actually occur. From this perspective, the reason why pause insertions increase intelligibility is that they serve to counteract erroneous information in the unnaturally compressed preceding part of the signal. This makes the relations between events in the signal more consistent with the temporal expectations speakers will have ordinarily acquired. In other words, the analysis presented above indicates that pauses – which are most frequently found at the boundaries of the syntactic and prosodic phrases that tend to prompt speech rate transitions (Grosjean et al., 1979; Miller et al., 1984) – may serve to help speakers reset and reorganize their expectations, making changes in the signal informative via systematic alterations to segmentation rate (or phase).

5.3 A theoretical account of the contribution of pauses to speech alignment

Our hypothesis is that pauses facilitate speaker alignment by making the rate at which informative events happen in the signal predictable and informative. Two simple, unambiguous predictions can be derived from it: first, if pauses serve as an alignment signal, then the distribution(s) of pauses in speech ought to be independent of speaker experience, such that they are identically distributed across speakers. Second, even when speaker expectations are aligned, variations in the rate at which informative events occur will need to be synchronized, and this will be reflected in the way pause distributions vary in time.

5.3.1 Formalizing a testable hypothesis

We next turn to determining the appropriate levels of analysis to apply to these predictions. First, how do we formalize and quantify both convergence and synchronization in the speech signal? We can summarize the requirements of alignment as follows:

1. the distribution of pauses ought to be structured so that speakers can extract information from the signal based on a relatively short exposure in time.

2. the information speakers extract from the signal should be independent of both the individual speakers' experience and fluctuations in pause durations and articulation rates throughout the interviews in the corpus.

In other words, **local** variation in pause duration when collapsed across speakers in time ought to be statistically independent of the history of the communicative

process, regardless of the timeframe used to operationalize history. If the distribution of pauses has this property, then, in theory, at any point in the communicative process, speakers' expectations about this aspect of the signal can be considered to be akin to a 'blank slate'. This critical property for alignment – that the distribution of some aspect of the time-varying signal be memoryless – has been shown to apply to discrete communicative events at various levels of description – i.e. lexical, sublexical, and phrasal – when they are considered in the communicative contexts in which they are used in text and speech (Ramscar, 2019; Linke and Ramscar, 2020b). That is, empirically, communicative events approximate geometric distributions at both higher and lower levels of analysis. This applies to linguistic events that are more discrete (e.g. words and phrases), and the more variant sublexical events they provide context for in speech. This finding is important because the geometric is the only discrete distribution that is memoryless.

Memorylessness describes a formal property of certain distributions of events where differences in knowledge about the events that have occurred prior to a certain point in time confer no advantage. This is because in these distributions, the variance in the rates at which events occur guarantees that knowledge about any events that have already occurred is uninformative in relation to predicting future events. Because the mean of the distribution is always equal to its variance, the variance is bounded by the mean, while the mean is bounded by the distribution of variance. The principle can be formalized in terms of a learning mechanism that seeks to minimize the variable error in the observations with respect to a stable average (cf. Peters and Adamou, 2022). When a distribution of events is memoryless, it follows that despite any local fluctuations, global event probabilities (or magnitude, in pause duration) are unaffected by knowledge of the history of the process. It thus follows theoretically, that once individuals have experienced a reasonable sample of such a distribution, their models of it will be largely independent of their idiosyncratic experiences. In other words, memoryless distributions ought to provide signals that allow for rapid adaptation of speaker expectations about the rate at which information will arrive.

It follows accordingly that if the distribution of speech pauses is memoryless, then it ought to approximate an exponential distribution (the exponential is the only memoryless distribution for continuous variables). Formally, the exponential is the probability distribution of the time intervals between events in a process in which events occur continuously and independently at a constant average rate. This means that once that speakers have learned to interpret them, local patterns of variation in signals will also have an experience-independent distributional structure. Thus, for present purposes the speech signal can be seen to provide two distinct sources of information: speech pauses which simply vary in duration; and articulated parts of signals that both vary in duration and frequency **and** which have their own

relational structure in sequences. The latter structure reflects the function of speech signals: that they are used to communicate, a process that critically relies on mutual predictability.

Formalizing context: defining a baseline to measure misalignment in time

The degree to which signals are predictable will be influenced by the amount of information available prior to a given point: their context. This raises a problem when we consider the communicative constraints speech codes must satisfy, namely that the experiences of the users of a code will vary greatly, both in what they have sampled and how often. This suggests that across speakers, individual ability to segment the signal and then use this to model context will vary. Moreover, these apparently separate processes will interact, such that as a result of this interaction both what a 'segment' is – and in particular, the amount of information required for its presence to be inferred – and the amount of information that 'segment' in turn contributes to contexts will both change with experience.

It follows that because articulations vary both with context and individuals' experience over time, they cannot themselves provide a baseline for our theoretical analysis of pause distributions. Rather, to test our hypothesis we require a signal dimension that develops linearly over time. One source of such information is sequence position. Sequence – or utterance – length is determined by variation in structural regularities at various levels, including word order, word morphology, and argument patterns. The rates at which regularities at these first two levels – which roughly correspond to syllable and short word boundaries – are produced appear to be relatively invariant both in terms of development of individual speakers and different languages (Poeppel and Assaneo, 2020; Luo and Poeppel, 2007). By contrast, sequence-level patterns of development at any timescale are characterized by distributional changes that involve an increase both in the length of sequences and the frequency of shorter sequences. Accordingly, we will use sequence position and its pattern of development with experience as a baseline in our analysis. We will examine the relationship between this baseline and our hypothesized time-invariant source of information provided by pauses and analyze the predicted changes this leads to with increasing speaker experience.

The hypothesis: predictable divergence between distinct signal dimensions over time

If pauses provide information about the rate at which other informative events occur in signals, and if experience changes the rate at which events occur (and hence

even their nature), it follows that the relationship between pauses and shared rates (i.e., word and syllable boundaries) ought to change at those points that are most affected by experience.

Lifelong experience is typically accompanied by some form of specialization. Adulthood typically involves the pursuit of different vocational paths and different interests. These inevitably increase both context-specific knowledge and vocabulary (Ramscar et al., 2014), which ought to be reflected in the production of signals that are increasingly context-specific, that we expect in turn to interact with sequence position. This prediction can be explained as follows: we have described above how communication of context-specific knowledge relies on – and hence will develop through – a transmission process that requires the maintenance of mutual predictability. We have suggested that this is achieved by the systematic adaptation of phrase structure and sequence length, a process that, at the utterance level, serves to reduce communicative uncertainty across structured sequences. Accordingly, it follows that if segmentation rate increases are a function of the relative predictability of the structured sequence, experience ought to lead to an increase in segmentation rate and a decrease in segmentation rate variability in later sequence positions.

A further implication of the interaction between experience, sequence position, and segmentation rate is that because context-specific signals are irregularly distributed themselves, experience will also lead to increases in the variability of local segmentation rates – bursts of activity – as opposed to global increases that would spread out uniformly across older speakers' signals. Notably, the correlation between information rates and signal sparsity appears to be an instance of a more general phenomenon. Bursty patterns of activity are a characteristic of complex event dynamics, such as for example traffic and internet activity (Karsai et al., 2018), and the way words occur in language use (Katz, 1996; Altmann et al., 2009). Which brings us back to the analogy between intervals between general classes of events and speech pauses. While gaps between events are a well-known quantity in the analysis of complex event dynamics, to our knowledge intervals between events have not previously been considered a source of information in relation to distributions of automatized complex cognitive behaviors. Yet, given that it has been shown that fluctuations in inter-event times are a better predictor of irregular bursts of activity characteristic to complex behaviors in humans than the event distributions themselves (Goh and Barabási, 2008), there is reason to suppose that they might also provide information about the processes that give rise to these behaviors.

As reviewed in section 2, pauses from different ranges of duration modulate the extent to which exposure to acoustic signals leads to temporal adaptation, and whether this exposure also results in adaptation to sensory prediction error. These two signal dimensions - timing and the information provided by the frequency spectrum

- contribute to distinct aspects of behavior: first, uncertainty management, which involves coordination of expectations in time; second, permanent adaptation in patterns of execution (e.g., the way speech signals are produced in context). Both of these mechanisms involve learning and take time to develop. The extent to which speakers' ability to utilize them efficiently changes with experience is determined by the structure of the speech samples they are exposed to, and the order in which distinct aspects of the signal become predictable and thus informative. At phrase level, for example, it is less likely that adult speakers will encounter novel verb argument structures. By contrast, the number of nouns they encounter within these structures will increase steadily across the lifespan (Ramscar et al., 2014). As a consequence, the information provided by verb argument structures will decrease in relation to the increasing number of context-specific nouns conditioned on them. More generally, experience will lead to an increasing asymmetry between variability (and uncertainty) in some aspects of the signal (e.g., context-specific articulations) in relation to the relative invariability of the communicative contexts they are conditioned on (Ramscar, 2021b). Simultaneously, more experienced speakers will acquire increasing certainty about those aspects of the signal/environment that are relatively invariant in time – e.g., how to structure sequences for successful uncertainty management and efficient message transmission.

These theoretic assumptions allow us to formalize two key questions we shall now address: First, are speech pauses distributed so as to allow speaker alignment in the way we suggest? Second, are the interactions between pause duration, timing, and speaker experience consistent with our predictions? If speech sequences are structured for incremental uncertainty reduction, the average information rate should increase with the relative sequence position (order). This means that information in speech ought not to be uniformly distributed in sequences (as indeed is the case (Genzel and Charniak, 2002; Genzel and Charniak, 2003)). We suggested above that this asymmetry in the temporal distribution of information in speech sequences will increase with speaker experience, increasing the differences between the rates at which individual speakers segment the signal. We hypothesize that speech pauses serve to offset these differences and help speakers organize their expectations about the rate at which information arrives in time. These considerations yield three concrete predictions that we will test in the analysis that follows:

1. That the aggregated samples of pause durations will converge on the exponential distribution.

2. That experience will optimize speech production by local adaptation of signal variability, which will lead to an increase in information density ('segmentation rate') in the utterance final positions of longer utterances. That, in final positions

of longer utterances, the uncertainty reduction by the previous part of the utterance allows information to be transmitted at higher rates, and a decrease in segmentation rate in the utterance initial positions and short utterances, where the uncertainty is high and information rates are low. This in turn will lead to an increase in signal sparsity and increasingly bursty patterns of activity in pause production.

3. That this tendency ought to be reflected in a systematic divergence from the mean pause duration with phrase length and the proximity of the phrase boundaries in speech produced by older speakers – i.e., we should expect more extreme values at the boundaries of both very short and very long phrases.

5.3.2 Corpus Data

We analyzed pauses from the Buckeye Corpus of Conversational English (Pitt et al., 2005) and the Korean Corpus of Spontaneous Speech (Yun et al., 2015). The corpora each contain phonetically transcribed speech from informal interviews with 40 speakers balanced by age and gender. The young cohort in the English corpus consists of speakers below the age of 30, the older speakers are aged 40 and upwards, with more age variance in the older speaker group. The Korean speakers are aged 15 to 47 years, with a median of 29.

All phrases produced by the interviewees were manually transcribed and word and phone boundaries were annotated with an automated speech aligner. Subsequently, the aligner annotations were manually corrected by trained annotators. Periods of silence that occurred between word boundaries were tagged as pauses. Absences of articulation accompanied by audible breathing or other sound were tagged separately. Short periods of silence that occurred between word boundaries were labeled as silent phones, stop closures, or assigned to surrounding phones (Kiesling et al., 2006). For the purposes of the analyses reported here, we focus exclusively on silent pauses between words.

The Buckeye corpus contains 26952 and the Korean corpus 21571 pauses. 98.7% of the pauses in the English sample and 99.75 % in the Korean sample are shorter than 3000 ms. To facilitate the analysis of pause duration in relation to utterance position, each pause was assigned to the utterance position of the word it preceded.

We examined the interaction of pause distribution, segmentation rate variability, and experience, using two parameters: utterance position and age cohort. Compared to the Korean sample, the English sample provides a larger span of speaker ages

and concomitantly speaker experience. This enables us to analyze speech data that allows for a strong test of this hypothesis. Previous findings have shown that response differences between age cohorts can be explained by the systematic changes of co-variation patterns in the samples speakers are exposed to and learn from across the lifespan and that the onset of these 'aging effects' is measurable in speakers' performance relatively early (in their twenties, see Figure 5.4, panel D) (Ramscar et al., 2013d). Therefore, despite the differences in the age range, we can reasonably expect to find measurable differences in the performance of the 'older' group of speakers (speakers above the age of 30) in the Korean sample.

5.4 Results

For the pauses extracted from the English and the Korean corpus we estimate the rate parameter λ, and model the exponential distribution using maximum likelihood estimation (for details, see Supplementary Materials).

Figure 5.1 shows the model and empirical distribution with best fits to the target range of durations (0-3000 ms). The empirical distribution of speech pauses in both English and Korean approximates the exponential, showing some divergence in the high-frequency range that hosts the shortest pauses ($< 100ms$) and in the tail of the distribution, where the random sample model fit predicts a larger range of values (extremely rare pauses longer than 3000 ms). The model distributions fitted to the selected range of values, such that the simulation approximates the best fit within the range (instead of the best fit to the number of events imposed by the corpus size), and the empirical distribution shows a close fit to the exponential model and the truncated random sample of exponentially distributed values.

To test for the distance between the model and empirical distribution we performed a two-sample Kolmogorov-Smirnov test. The test shows a relatively small, but significant difference between the empirical distribution and the truncated model, in both languages ($D_{ks} = 0.09, p < 0$). It seems worth noting here that the test is very sensitive to small differences between distributions and also shows a significant difference between the truncated exponential and the exponential model ($D_{ks} = 0.01, p < 0$). The distance between the empirical distribution and the model reflects the misalignment in the high-frequency part of the distribution that holds the shortest pauses. This is unsurprising given that unvoiced stops and silent vowels (whose durations vary in the range of 20-100 ms) were excluded from the analysis. (Detailed test statistics and discussion of alternative distributional hypotheses are provided in the supplementary materials.)

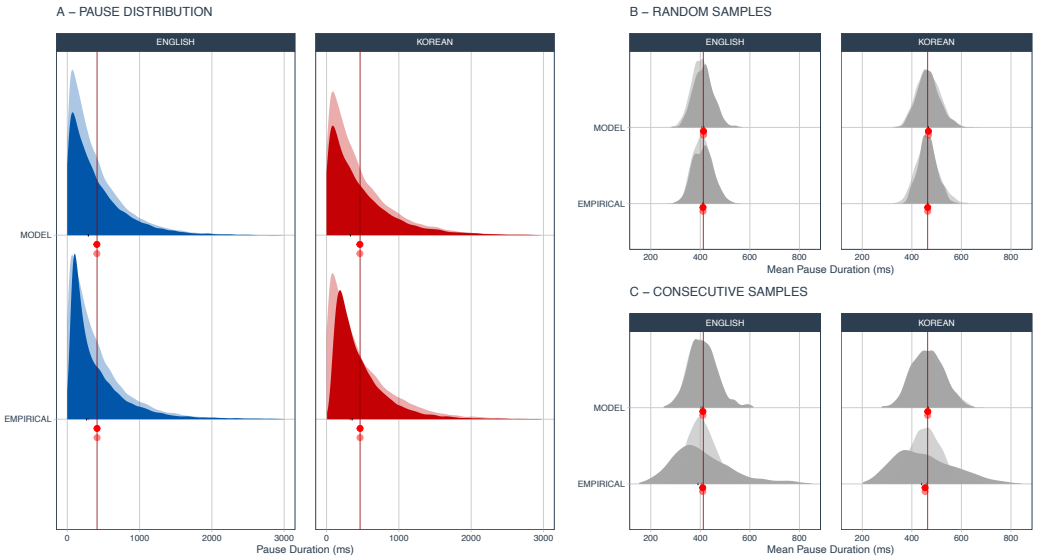

Fig. 5.1: Left panel (**A**): probability density function of pause distribution for pauses shorter than 3000 ms (bottom row) and a random sample from an exponential distribution (top row) of identical size and rate parameter. The shaded area shows the truncated exponential model, limited in the analyzed range of values $(0, 3000)$. The sample means (red dot) and model mean (transparent red dot) center at the theoretical mean (red line). Right panel: the distribution of sample means in samples ($n = 50$) drawn from the distributions shown on the left. Means from random samples (**B**) converge on a normal distribution that centers at the theoretical mean. Means from samples of consecutive events (**C**) show the same behavior for model distributions, while means from consecutive samples drawn from the empirical data are both more dispersed and left-skewed, indicating local biases in the distribution of pause durations.

Consistent with previous findings (Linke and Ramscar, 2020b), individual speakers' samples seem to deviate randomly, with samples approximating the gamma distribution with varying scale parameters. The aggregate approaches the exponential distribution (which is a special case of gamma). This means that despite a multitude of sources that concomitantly feed into individual variation (i.e. individual experience, syntactic structure, and various other factors that shape interactions and rates at which acoustic events are extracted from signals) the aggregate distribution maintains a structure that can allow speakers to converge on a stable mean expectation. To examine the effects this individual variation has on the local distribution of variance, we run a sampling simulation. In order to investigate whether speech pauses exhibit bursty patterns of activity, we draw 1000 random samples of 50 'pauses' from each distribution and samples of 50 consecutive 'pauses', comparing the distribution of mean duration in relation to the theoretical mean $1/\lambda$. The right panel shows the distribution of averages from random samples (top) and sequences (bottom) for Korean (right column) and English (left column). Sample averages from random samples and samples of consecutive events from the model distribution converge on a Gaussian distribution that centers at the theoretical mean. We observe more variance in the means of sequential samples. Samples of consecutive events

	RANDOM					
ENGLISH	Mean	Median	SD	SE	IOD[1]	CV[2](%)
model	407.63	407.35	60.26	1.91	8.91	14.78
empirical	409.96	405.86	59.85	1.89	8.74	14.60
KOREAN						
model	468.13	463.06	67.96	2.15	9.87	14.51
empirical	464.14	461.25	51.89	1.64	**5.80**	11.18

	CONSECUTIVE					
ENGLISH	Mean	Median	SD	SE	IOD	CV(%)
model	412.40	409.51	54.77	1.73	7.27	13.28
empirical	412.12	390.90	126.83	4.01	**39.03**	30.78
KOREAN						
model	466.75	464.31	61.56	1.95	8.12	13.19
empirical	468.23	447.11	130.32	4.12	**36.27**	27.83

[1]Index of Dispersion, [2]Coefficient of Variance (%)

Tab. 5.1: Distribution of pause duration from samples of conversational English and Korean. Mean durations in random samples drawn from the empirical distribution of Korean pauses are less dispersed than the model mean distribution. As shown in Figure 1 of the results section of our article, samples of consecutive pauses from the empirical distribution are more dispersed than consecutive samples from the model distribution. This indicates that pauses from different ranges of duration are not uniformly distributed in time, and instead appear in local bursts.

drawn from the empirical distribution, show a strong left skew and more dispersion, suggesting that extreme values in pause durations are not distributed uniformly in speech sequences.

These results suggest that pause distributions possess two desirable properties: one, a time-invariant global distribution that can provide developing speakers with a consistent source of information in the otherwise noisy signal they are exposed to; two, systematic local fluctuations in average pause duration that can allow adult speakers at different levels of experience to rapidly adapt to changes in local segmentation rates given relatively short signal samples. All of which raises a critical question: are local fluctuations in mean pause durations systematic in relation to changes in the distribution of information (and uncertainty) in signals?

How does pause duration interact with the event rate? What is the event rate?

Formally, the exponential is a distribution of waiting times between events generated by a Poisson process. The λ parameter defines the rate at which these 'events' occur: the average number of 'events' within a fixed time interval. We discussed earlier how in language, the informative changes in signal dimensions that can constitute informative 'event' boundaries shift more or less systematically with speaker experience. As a consequence, the rates at which the respective 'events' occur ought to shift more or less systematically too. It follows therefore that rate estimates only ever approximate the convergence rate, and do not represent fixed, objective estimates of the rate at which individuals will extract informative changes from the signal.

With respect to the pause distributions, which as we have shown in both Korean and English approximate the exponential distribution, the rate parameters ($\lambda_{ko} = 0.0022$, $\lambda_{en} = 0.0024$) indicate that within a fixed frame of 1000 ms, Korean speakers ought to converge on a 'minimal agreement' that informative changes in the signal can be 'expected' to occur every 464 ms ($1/\lambda$) on average and that, given that pause likelihood decreases exponentially with pause duration, the probability of subsequent acoustic events occurring increases exponentially with pause duration. In other words, if pauses are information and information processing is probabilistic, an exponential decrease in probability as a function of pause duration indicates that there are large differences in the way small perturbations in pauses from different ranges of duration interact with probability.

In this model, the probability of an informative change (articulation) increases considerably as pause duration increases from 100 ms to 150 ms, while hearing more silence is almost equally unexpected between 750 to 800 ms. This in turn provides an interesting probabilistic perspective on the possible role of pause duration in the dynamics of uncertainty management, and the response to informative changes in articulations. It suggests that all articulations following longer pauses are uninformative because they are over-expected, while the informativeness of articulations following short pauses will vary with the expectations built up by the previous part of the articulated signal. If we apply the same reasoning to pauses from the middle range, then the information provided by subsequent articulations ought to be a compromise function of both pause probability and the articulated prior. With respect to pauses in speech, and the way they manage expectations (i.e., the information rate), an over-expected stretch of speech signal immediately following a long pause can be conceptualized as follows:

> [1269 ms PAUSE] *that* a president would do something ...

In this example, the pause duration leads to an expectation of the way *that* will be signaled. In order to be intelligible in isolation, *that* would ordinarily be signaled using multiple, discriminable acoustic segments *dh ae t*. However, in the context of the long pause that precedes it, the measurable acoustic-phonetic information it contains can deviate from the isolated signal as long as the information it contributes does not violate speakers' expectations about what **can** follow. That is, at points where articulated signals are over-expected, the only thing that will violate expectations is the absence of articulation – hearing any articulation is more probable than hearing more silence. Simultaneously, speakers' expectations about the way a signal following a pause will unfold in time (where word and syllable boundaries occur) appear to be relatively fixed. This means that whatever is articulated instead of *dh ae t* – even if it is a mere burst of noise – must occur within the expected time interval in order to maintain the intelligibility of those parts of the signal that follow it (see Doelling et al., 2014).

At the other end of the pause spectrum, the fine-grained acoustic detail that follows the bulk of short pauses ought to become increasingly informative and diversified in signals in which they occur. Because short pauses provide little useful information on their own, listeners ought to get better at noticing whether any violation of prior expectation actually occurred in the acoustic signal and respond with fine-grain temporal adaptation. From this perspective, the articulated signals that are learned and remembered are systematic independent of speaker experience, because they are shaped by the pause distribution and vary consistently with pause duration.

Accordingly, it follows that if the exponential model applies to speech pause distributions in the way we suggest, then segmentation rates ought to become increasingly diversified with experience. To investigate this question we analyze experience-related change in utterance structure and the distribution of pauses.

5.4.1 The interaction between sequence length, pause duration, and experience

We analyze the frequency distribution of utterance positions and pause durations. Prior analyses have shown that the distribution of words across utterance position, and word length in phonemes in speech data follows the geometric distribution. This indicates that segmentation rates that approximate word and phoneme boundaries in spontaneously produced signals vary consistently in time. The correlation between log-transformed frequency and frequency rank in these distributions reflects their fit to the geometric distribution.

The analyses presented in Figure 5.2 show that frequency of utterance positions closely approximate a geometric distribution in both languages and cohorts ($R^2_{en} > 0.994$, $R^2_{ko} > 0.997$). While Korean utterances are on average shorter ($M = 4.017$, $SD = 3.302$, $n = 57661$) than English utterances ($M = 5.710$, $SD = 5.357$, $n = 50841$), older speakers of both languages produce longer utterances more often (Mann-Whitney, Korean: $M_o = 4.37, M_y = 3.74, U = 0.47, p < .0032$, English: $M_o = 6.13, M_y = 5.50, U = 0.28, p < .0763$). The medians are significantly different in older speakers of English (Mood, Korean: $Mdn_o = 4.14, Mdn_y = 3.64, Z = -1.13, p - .2573$, English: $Mdn_o = 6.20, Mdn_y = 5.65, Z = 2.43, p < .015$), indicating divergence in scale and more individual variability in utterance length in older speakers of English. Note that we are not interested in individual variation in the analyses reported here, rather the analysis explicitly targets aggregate behavior. The results show that the older speakers maintain the optimal distribution **despite** the increase in variability across individuals. (All descriptive statistics and results of statistical tests are provided in the supplementary materials.)

As discussed above, spontaneous behavior can be expected to vary across individuals, and with experience, the variance in the articulated parts of signals can be expected to increase. Accordingly, we examined whether the variance in individual behavior does in fact systematically increase in the aggregate (despite the many sources of variation that shape individual behavior in seemingly unsystematic ways). As expected, the distributions maintain their shape (they stay geometric) and as a result, the slope of the older speakers' distribution decreases. The changes in slope indicate a decrease in the information provided by the relationships between words, i.e., their relative probabilities in context. Typically, decreases in performance variability over a finite set of outcomes like this are taken as a mark of learning, and the decrease in noise to signal ratio associated with it (cf. Shmuelof and Krakauer, 2014; Tucha and Lange, 2004; Herman et al., 2009).

This course of development is consistent with the idea that changes in the signals produced by older speakers reflect the patterns of learning that can be expected to accompany increased experience rather than any 'pathological' decline in performance (cf. Ramscar et al., 2014; Campbell et al., 2016). The decrease in structural uncertainty (i.e., the amount of variance in the conditional probabilities among adjacent words) in older speakers facilitates a redistribution of functional load: the uncertainty at word transitions decreases, freeing up resources to allow learning of fine-grained, context-specific patterns of articulation (cf. Poulisse et al., 2020).

However, whereas we observe a marked effect of experience on the articulated part of signals, it does not appear to affect pause distributions. The frequency distributions of pauses (with pause duration binned at 10, 20, or 50 ms) in both age groups closely approximate the geometric ($R^2_{en} > 0.984$, $R^2_{ko} > 0.993$) and

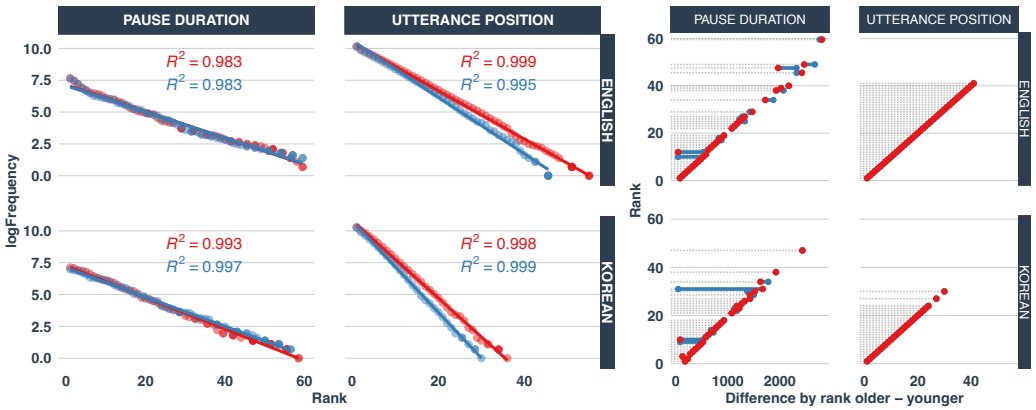

Fig. 5.2: Left panel: Frequency distributions of utterance position (left) and binned speech pause duration (right, bin size 50 ms) for younger (blue) and older (red) speakers. Lines show model fits to geometric (linear after half log transformation). Both the pause distribution and the utterance position distribution show close fits to geometric with and $R^2 > 0.99$respectively. Note that the lines that represent the geometric model distribution are hard to distinguish because the distribution and the model fits are nearly identical. In the left column, the younger and the older cohorts' distribution and their respective fits overlap completely. Right panel: pointwise correlations of pause duration (left) ranked by frequency show some minimal misalignment in the rank distribution. As can be seen in the rightmost column, the ranked distributions of utterance positions are identical. Older speakers produce longer utterances on average while maintaining the shape of the distribution and the relationship between utterance length and utterance probability. This correlation is unusual, when same analyses are performed on text, the correlation between ranked values from any two samples in a mixed corpus tend to be weaker.

both distributions have identical slopes. Simultaneously, the mean pause duration decreases in the older cohorts. The differences in the means are not significant (Mann-Whitney, Korean: $M_o = 439.28, M_y = 454.96, U = 0.1, p = .5291$, English: $M_o = 400.88, M_y = 408.21, U = 0.0698, p = .6588$), and neither are the differences in the medians (Mood, Korean: $Mdn_o = 419, Mdn_y = 441, Z = -0.95, p = .3398$, English: $Mdn_o = 381, Mdn_y = 413, Z = -0.12, p = .9081$). In English, we observe more variability across older speakers and a decrease in variance in the aggregate. (See Supplementary materials for a full summary of descriptive statistics.) A closer inspection of differences in the probability density (Figure 5.3) reveals that older speakers of both languages produce more pauses from the middle range of durations (250-750 ms). In other words, older speakers of both languages produce fewer pauses that markedly diverge from the mean duration. This suggests that in pause production, experience leads to a gradual convergence on the mean.

To summarize these findings, experience appears to result in a redistribution of uncertainty across utterance positions as it increases. The differences between the distributions indicate that experience-related changes in the distribution of uncertainty in the articulated parts of signals increase linearly with utterance position. The fact that the distribution of pause durations does not change with experience

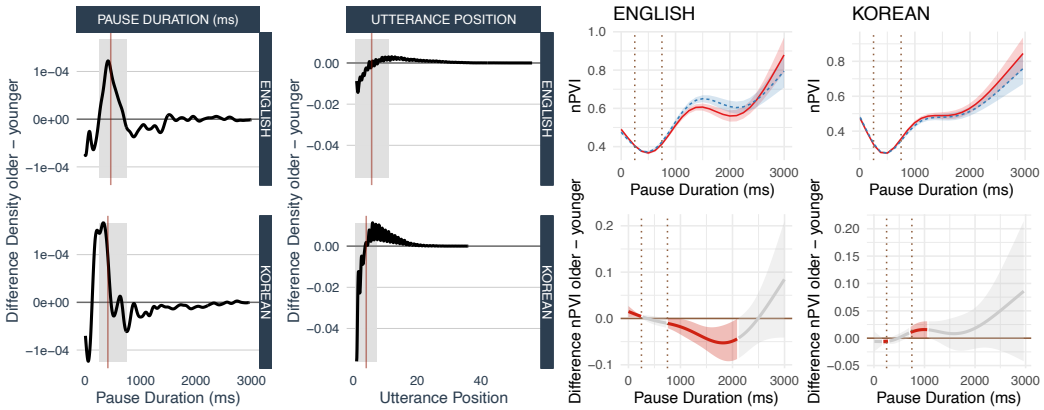

Fig. 5.3: Left panel: Difference in probability density of utterance position (right panel) and pause duration (left panel) in speech samples produced by older speakers of English (top row) and Korean (bottom row). The red line marks the sample mean, the gray area highlights pauses from the middle range of durations (250-750 ms) and the standard deviation from the mean in the probability of utterance position. As can be seen, plots reveal a shift towards the mean (convergence) in pauses shorter than 1000 ms, indicating that longer than average pauses become shorter and shorter than average pauses become longer in older speakers, decreasing the individual variance in pause duration. By contrast, for utterance position experience shifts the average towards the right showing an increase in variability (i.e, while pauses converge, utterances appear to diverge). The effect is more protracted across utterance positions in English speakers, which is consistent with the longer average utterance length in English. Right panel: Pairwise pause variability (**nPVI**) as a function of pause duration and speaker age. The top panel shows the relationship between pause duration and pairwise variability, which is identically u-shaped in both languages, reaching its minimum at the mean pause duration. The bottom right panel shows the differences between cohorts (areas, where the difference is significantly larger than 0, are highlighted in red), which show opposite patterns in Korean and English.

indicates that the relationship between pause duration, utterance length, and uncertainty will change systematically over time. The slopes in the distribution of utterance positions decrease, indicating a decrease in differences in the distribution of uncertainty across utterances. This implies that if pause durations reflect uncertainty, the relative differences between consecutive pauses should also decrease with utterance length. This also suggests that alignment in communication relies on two distinct notions of convergence in probability. Both alignment and stochastic convergence rest on the idea that a stream of random or unpredictable events or quantities (noise) can settle into a predictable behavior over time. A predictable behavior can be characterized by:

1. a decrease in contrasts between consecutive values (the observed events or quantities eventually become indistinguishable)

2. a stable probability distribution, where the contrast between consecutive values is maintained, but the change is kept predictable (which can be achieved by learning to ignore unpredictable variation)

We propose that communication between speakers at different levels of experience relies on both of these different notions of convergence and that the degree to which speakers utilize one or the other changes across the lifespan. So far, we have described these distinct aspects of alignment as structural or systemic (pertaining to relationships between informative events) and contrastive (pertaining to discriminable changes in the informative events themselves). Consistent with this theoretical analysis, these results suggest that experience acts as a 'sink' in that it decreases the variability of higher-level structural or relational information, eventually making highly predictable events indistinguishable.

To further examine this, we next analyzed the effect of experience on the relationships between consecutive pauses and their interaction with pause duration. Changes in the relationships between consecutive pauses indicate uncertainty changes and the rate at which the signal is segmented. Larger differences ought to indicate larger changes in segmentation rates of consecutive articulation. That is the deviation from the 'event' or segmentation rate in the articulated stretch of signal that separates two pauses ought to be reflected in the durational contrast, such that convergence in rate, or an increasingly uniform distribution of information, mediated by the convergence in sequential order, would result in a decrease in durational contrast between consecutive pauses.

5.4.2 Changes in durational contrast between consecutive pauses

To quantify these changes in durational contrasts, we calculate the normalized pairwise variability index (**nPVI**) of adjacent pauses (pairs of pauses separated by an articulation). This metric quantifies the average durational contrast between consecutive events in acoustic sequences (e.g., notes in a melody or vowels in speech utterances). It was originally developed to distinguish timing patterns in languages that differ in prosodic structure (Ling et al., 2000; Grabe and Low, 2008) and captures the relative variability of segment durations in relation to the event onsets of isochronous sequences. The nPVI for a pause occurring at position *k* and *k+1* in a sample of *m* silent pauses of duration *d* is calculated as

$$\sum_{k=1}^{m-1} \frac{|d_k - d_{k+1}|}{(d_k + d_{k+1})/2} / (m-1)$$

The metric allows us to estimate differences in the duration of adjacent pauses independent of local fluctuations in the articulation rate. The nPVI ought to reveal changes in the temporal relationships independent of the pause durations themselves. i.e., if pauses do inform timing, relative changes in duration provide information about changes in the segmentation rates of articulated signals that separate two consecutive pauses, independent of context-related variability of segmentation rates.

We observe a decrease in the average nPVI in older speakers of both languages (Mann-Whitney, Korean: $M_o = .336, M_y = .345, U = .154, p = .327$, English: $M_o = .425, M_y = .430, U = 1.462, p = .713$), which could indicate a decrease in rate variability in the articulated parts of the signal that constitute the transitions between consecutive pauses. Moreover, there is less variability in temporal contrast between adjacent pauses in Korean (Mann-Whitney: $M_{ko} = .341, M_{en} = .427, U = 1.462, p < 0$), a trend similar to those previously observed in vowel durations (Grabe and Low, 2008).

We model the relationship between pause duration and the nPVI as a non-linear two-way interaction with experience (Figure 5.3, left panel) in a generalized additive model using the R library *mgcv*. Generalized additive models allow us to examine non-linear interactions between multiple predictor variables while accounting for the highly skewed distribution. We fit a Gamma model with a log link function. The model is specified as follows:

$$nPVI \sim s(PauseDuration, by = cohort, k = 5)$$

We find a u-shaped effect of pause duration on durational contrast in English (older: $edf = 3.977$, $F = 187.6$, $p < 0$, younger: $edf = 3.976$, $F = 245.9$, $p < 0$) and Korean (older: $edf = 3.978$, $F = 207.8$, $p < 0$, younger: $edf = 3.976$, $F = 176.0$, $p < 0$). To test for differences between the cohorts, we examine the difference curve between the smooths of the two factor levels (older and younger speakers). The relationship between pause duration and durational contrast to adjacent pauses does not differ significantly between the cohorts in pauses from the middle range of durations, whereas cohort differences in pauses shorter than 250 ms and longer than 750 ms are significant (Figure 5.3, the bottom left panel shows the difference between the smooths of the two cohorts, the factor levels 'older' and 'younger', parts of confidence intervals that do not include 0 are highlighted in red). In pauses produced by older Korean speakers, there is an increase in durational contrast between shorter pauses and their neighbors, and a decrease in longer pauses. By contrast, older speakers of English decrease the durational contrast between longer pauses and successors and increase it in shorter pauses.

To test whether the differences in the duration of pauses produced by older speakers is correlated to the increasing sparsity in the distribution of information in the way we suggest, we next model the probability and duration of pauses produced by younger and older speakers as a function of utterance length and utterance position.

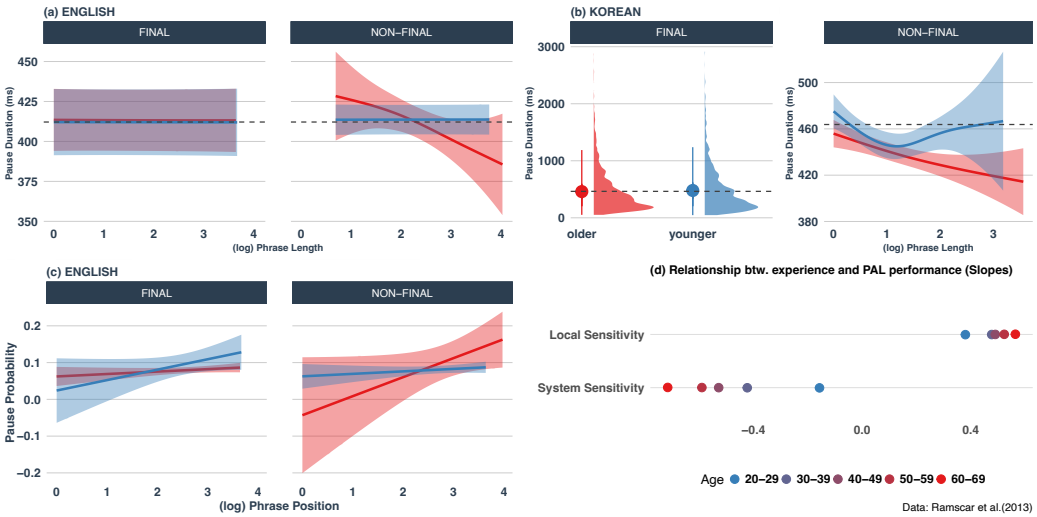

Fig. 5.4: Mean pause duration as a function of utterance length for older(red) and younger (blue) speakers of English (a, top left) and Korean (b, top right). In the Korean sample, pauses occur at phrase initial boundaries only, pauses in the English sample are distributed across the utterance - the distribution of utterance position of words preceded by pauses is geometric. Pause probability decreases with utterance position in both younger and older speakers, but pause probability appears to shift towards the later positions in older speakers (panel c): In English, experience increases the average duration of pauses preceding non-final words in shorter sequences and decreases the average pause duration in longer sequences. In Korean, the cohorts diverge in pause durations preceding longer utterances only but generally appear to model a similar behavior, which becomes less noisy with experience – older speakers produce shorter pauses, and there is less variance in duration. **Panel d** shows the model expectation: effect of experience-related changes in English speakers' sensitivity to the frequency of co-occurrence relationship (system sensitivity) and cue frequency (local sensitivity) on performance on the paired associate learning task by age group.

5.4.3 Does experience change the relationship between pause and utterance length?

In the English sample, pauses are distributed across utterances and the distributions of utterance positions preceded by pauses are geometric; they do not differ from the overall distribution in both cohorts. In Korean, silent pauses are found at the utterance initial boundary only. This is consistent with the differences in the prosodic structure of Korean, which tends towards isochronous syllable timing (Tark, 2012; Moon-Hwan, 2004), and English, which is stress-timed and highly variable in syllable timing.

We fit a generalized additive model of pause duration as a smooth function of utterance length for each cohort, adding a factor term for the utterance final boundary (a categorial predictor to distinguish between final and non-final positions of utterances). We use a Gamma model with a log link function to account for the highly skewed distribution. The model is specified as follows:

$$PauseDuration \sim s(UtteranceLength, by = cohort) + UtteranceFinal$$

The model estimate of the average duration of pauses in non-final positions is 411 ms in English and 444 ms in Korean. Only older speakers of English seem to deviate from the average pause duration in non-final positions of longer utterances. In the model, the average pause duration in older speakers of English decreases linearly with utterance length (English, older: $edf = 1, F = 6.5620, p = .0104$, younger: $edf = 1, F = 0.031, p = .8610$, Korean: older: $edf = 1, F = 1.028, p = .311$, younger: $edf = 1, F = 0.429, p = .513$).

The Korean data does not support our hypothesis that experience ought to increase the burstiness of the pause distribution in older speakers. On one hand, this could be an effect of the relatively large differences in the age range of the English and Korean cohorts. On the other, as mentioned above, the temporal structure of Korean speech sequences differs fundamentally from the temporal structure of English. Alternation in syllable duration is an important functional feature of the sequential structure in English: the variability in stress patterns is a source of information. Korean, by contrast, is markedly less variable in vowel (and syllable) duration and appears to approach an isochronous temporal structure similar to mora-timing in Japanese (Tark, 2012; Moon-Hwan, 2004). This suggests that the functional load on alignment in timing might be substantially reduced and more evenly distributed across other functional features of the signal in Korean. To explore whether the differences in the cohort effects reflect a general difference in the distribution of pauses and durational contrasts in English and Korean, we conducted a final set of sampling simulations to address this question.

5.4.4 Contrasting age-cohort differences in pause variability and pause duration in consecutive and random samples

To examine the distribution of sample averages we obtained means from blocks of consecutive pause durations and normalized durational contrasts (**nPVI**) from older and younger speakers' speech. We simultaneously extracted means from blocks of randomly ordered pauses to serve as a baseline. Figure 5.5 shows the distributions of mean pause duration and mean nPVI for blocks of 20 pauses. Pause

duration averages from random samples approximate a Gaussian distribution that centers around the theoretical mean in Korean and English. The mean nPVI in random samples is higher than the mean nPVI in consecutive pauses in English ($M_{ran} = 0.4362, M_{cons} = 0.4267, U = 0.2564, p < 0.01$). The difference between the model nPVI and the empirical nPVI indicates that the order in which pauses from different ranges of duration occur decreases the average distance between adjacent pauses in English, but not in Korean ($M_{ran} = 0.3595, M_{cons} = 0.3572, U = 0.0815, p = .15$).

In the cohort comparison, the distribution of nPVI averages from samples of consecutive pauses indicates local reductions of durational contrast in older speakers of English ($M_o = 0.417, M_y = 0.435, U = 0.1958, p < .001$) and Korean ($M_o = 0.34, M_y = 0.378, U = 0.2564, p < .001$). There is a significant difference between the random and the empirical distribution of pairwise variability in English ($M_{ran} = 0.436, M_{cons} = 0.427, U = 0.147, p < .01$). We find no such difference in Korean ($M_{ran} = 0.359, M_{cons} = 0.357, U = 0.081, p = .15$). This 'clumpiness' in the relationship between consecutive pauses in English suggests that the order in which temporal events unfold is more informative in English than in Korean. This is consistent with findings that reveal differences in the extent to which language users exhibit preferences in perceptual grouping of rhythmical sequences. English speakers are sensitive to the relationships between durations of consecutive events, whereas speakers of the more rhythmically consistent Japanese do not exhibit strong grouping preferences (Iversen et al., 2008). In other words, temporal fluctuations of consecutive intervals constitute signals in English. This does not appear to be the case in Korean (or Japanese).

The distribution of pause durations from consecutive samples also reveals differences between the languages. In English, distributions of averages are bimodal (suggesting local bursts of activity) and show diverging patterns in older and younger speakers: bursts of shorter than average pauses in older speakers and longer than average pauses in younger speakers. This tendency increases in the later interview blocks, suggesting that the effect is cumulative and develops at multiple timescales (utterance, interview time, and speaker experience). In Korean, we find the opposite effect: older speakers' averages are markedly less dispersed and show more convergence across local samples, which can mean that the variability in pause duration decreases globally in Korean. Younger Korean speakers' samples appear to be far more variable in average duration and look more similar to older English speakers' distribution. The dispersion in consecutive samples we see in Figure 5.1 and table 8.1 appear to reflect bursty patterns produced by older speakers of English and younger speakers of Korean. We interpret these sample differences as divergent effects of lifelong experience on the distribution of pauses in spontaneous Korean and English. In English, differences between older and younger speakers increase

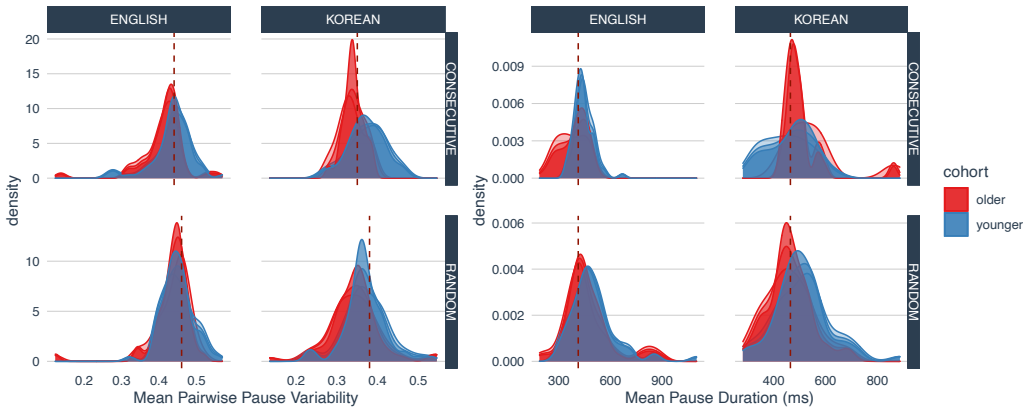

Fig. 5.5: Distribution of mean nPVI (left panel) and mean pause duration (right panel) in blocks of consecutive (top row) or randomly sampled events (bottom row) for older (red) and younger (blue) speakers. The dashed red line marks the sample mean. Mean values from random samples converge on a normal distribution that centers on the sample mean. Means from samples of consecutive pauses suggest differences between the cohorts and the languages. In Korean, there is more dispersion in younger than in older speakers' consecutive samples, which could indicate that local and global patterns of pause distribution get more similar (less sparse/bursty) across the lifespan. By contrast, the distribution of averages from the English sample turns increasingly bimodal with speaker age and interview time in consecutive samples, indicating an increase in local bursts of shorter pauses in older speakers and longer pauses in younger speakers that also increases throughout the interview (shaded areas show the distribution of samples from later blocks). The divergent patterns between the languages suggest that alignment is achieved through local optimization (bursts of activity) in English and globally (through an increasingly uniform distribution) in Korean.

at those points where experience-driven changes are greatest, at the boundaries of longer phrases. The results of the analyses indicate that patterns of pause production systematically interact with *local* changes in information rates, providing support for our hypothesis. In Korean, contrary to our prediction, experience appears to lead to a **global** decrease in the temporal variability of pause durations. Given the many uncertainties that follow from the differences between the languages and the speech samples at our disposal (the corpora are similar in many ways, yet not strictly parallel), however, these findings cannot serve as strong evidence of functional interactions between pause and information rate. Instead, they can serve to inform future research.

5.5 Summary and Discussion

In this article, we examined the hypothesis that speech pauses play a crucial role in the systematic temporal structuring of speech signals and the maintenance of mutual predictability in time. We presented a theoretical rationale for the hypothesis and evidence to support it. We suggest that pauses provide information about

fluctuations in articulation rates and that predictable interactions between pause duration and information rate facilitate alignment between speakers at different levels of experience. We approach the problem of alignment as a task of learning models that maintain mutual predictability of signals transmitted across multiple timescales. These models are learned from exposure to a noisy signal and structured by generations of speakers. We argue that a time-invariant source of information is necessary to counter the inevitable misalignment of expectations (i.e., differences in experience) that this complex distributed learning entails. We highlight how the distribution of information in human vocal signals contributes to learnability and efficient transmission across the lifespan.

We proposed memorylessness as a key prerequisite for successful alignment and efficient transmission and our finding that pause distributions from both speech sources, Korean and English, closely approximate the exponential distribution thus appears to fulfill this requirement. We hypothesized that to enable alignment in the way suggested, the exponential source ought to exhibit the following three properties:

1. a global distribution that provides developing speakers with a consistent source of information in the noisy signal they are exposed to

2. a time-invariant source of information to ensure that all speakers will acquire and maintain sufficiently similar models of expectations independent of the points in time they enter the speaker community

3. systematic local fluctuations in the temporal relationships between pauses to allow adult speakers at different levels of experience to rapidly adapt to local changes in segmentation rates

Our results indicate that pauses meet all three requirements. Their empirical distributions increase the sampling efficiency in random samples, and their global distribution is characterized by local fluctuations, which lead to a skew in the distribution of sample averages from sequences of consecutive pauses. To investigate whether these local fluctuations are systematically correlated to variation in the distribution of information, we analyzed the distribution of pauses across sequences of different lengths, comparing speech samples of younger and older adults.

We found that experience increases the average utterance length in conversational speech and that this change leads to a systematic redistribution of information across utterance positions. We suggested that this redistribution of information reflects a decrease in uncertainty about structural regularities that set the utterance context

and increase the efficiency of message transmission. The result is a difference in the distribution of information between the cohorts that increases linearly as a function of utterance position. The distribution of pauses, by contrast, does not change significantly with experience. The only change we observe in the distribution of pauses is a gradual reduction in the frequency of pauses that diverge notably from the mean pause duration indicating a global reduction of temporal contrast in pauses from the middle range of durations (250 - 750 ms) and a regression to the mean. Notably, we observe almost identical patterns of change in the cumulative distributions of English and Korean. Further, these effects are consistent with the predictions of the theoretical model derived from the quantitative structure of languages and learning theory.

Our results show that the misalignment in the distribution of information across utterances in relation to the relatively stationary distribution of pauses leads to a predictable decrease in the average pause duration in longer utterances produced by older speakers. This finding indicates a shift in the relationship between pause duration and the utterance position that is systematic – it scales linearly with utterance position – and is thus consistent with the suggestion that the relationship between pause duration and the changes in time intervals at which informative changes in the signal are observed are predictable.

Importantly, we find different patterns of age-related changes in the spread of Korean and English pause distributions. We hypothesized that the rates at which speakers segment the signal across sequences would become more variable with experience, increasing the signal sparsity (burstiness) over time. The effect is evident in the English sample, but not in the Korean sample. Contrary to our hypothesis, older speakers of Korean appear to minimize durational contrasts globally. By contrast, English speakers appear to minimize durational contrasts locally as predicted. Retrospectively, given that our hypothesis was that pause durations provide an alignment signal through consistent interaction with segmentation rate, the divergence in the patterns of lifelong development between the two languages need not be surprising. Our hypothesis was informed by evidence of the cumulative effect of experience on the performance of English speakers, consistent with the vocabulary development specific to the relatively impoverished information structure of English morpho-syntax (Ramscar et al., 2014). However, Korean and English differ significantly in the extent to which patterns of lexical productivity are implicit (characterized by regularities such as a rich morphology) or explicit (characterized by less productive, explicitly lexicalized forms) (Ramscar, 2021a). Korean relies heavily on a rich morphological structure to maintain systematic variation in the distribution of word forms. English morphology, by contrast, is far less informative such that far more meanings are realized in explicitly lexicalized, often idiosyncratic forms. From a learning perspective, this has a differential impact on lifelong development of vocabularies and the

distribution of functional load across speech sequences (contexts) in Korean and English.

One possible explanation of the effect is as follows: Language learning relies on regularities, such as for example, relative invariance in the patterns of inflection (Ramscar et al., 2013e). Unattested forms, e.g., regular plurals, can often be easily inferred from the general pattern of inflection (squid-squids, octopus-octopuses, wug-wugs, niz-nizzes), providing language learners with an important source of predictable variance. By contrast, irregular forms can only be learned explicitly (see Ramscar, 2021a, for review). By implication, the fact that the relatively impoverished morphological structure of English relies more on word forms that cannot be derived implicitly leads to a redistribution of functional load to those aspects of the signal that support the learning of explicit forms. Conversely, in Korean the functional load ought to be distributed across a variety of morphological cues (particles, endings and functional prefixes).

We have argued that the learning of explicit word forms and the alignment of speaker expectations in a communicative system characterized by non-linearity are facilitated by context. Because context is set both by the highly variable aspects of communicative codes such as regular patterns of co-occurrence between words, phrases, and segments (Ramscar, 2019; Linke and Ramscar, 2020b), but also the less variable aspects of speech signals such as prosodic stress patterns and the variability in duration (McQueen and Dilley, 2020), the differences in the lexical productivity of word forms we have described can result in very different predictions when it comes to the dynamics of lifelong learning. It is thus not entirely surprising that the experience-related changes in the patterns of pause distribution lead to increasingly uniform signal distributions in Korean and increasingly sparse signal distributions in English. We suggest that these differences in pause distributions could reflect differences in the dynamics of lifelong learning that interact with the linguistic structure set by the morpho-syntax and prosody/timing. These considerations highlight the necessity of extending these analyses to other languages that are placed along what seems to be a gradient scale of structural organization in human communicative codes.

6

Does the realization of spoken word morphology reflect speaker uncertainty? Speaker experience shapes speech form production across adulthood

In the previous chapter, we suggested that interactions between the utterance structure and the uncertainty related to the distribution of words in utterances affect speech form realization. As speakers' uncertainty about the structure of speech sequences decreases, the distribution of information appears to shift: words/signals in earlier utterance positions become less informative [1], and words/signals in later utterance positions become more informative [2]. This asymmetry in the distribution of information across sequences, in turn, seems to change the rates and timescales at which signals resolve. Findings presented in chapter 5 suggest that speaker experience systematically alters information rates across utterance positions and across utterances of different lengths. These observations raise new questions: does the redistribution of information change the way older speakers articulate words? Where can we expect such changes in signals to occur?

This chapter examines the effects of lifelong learning on the way words are articulated in context. In what follows, we show how adult experience shapes segmental variation in conversational speech. Taking regular patterns of co-occurrence between speech forms as a context for speech production and learning, we explore how the structure of cumulative samples can change the associative relationships between word pairs from different lexical categories and, with that, the relationships between words and contexts they occur in. Specifically, we examine how systematic changes in the structure of the speech samples speakers are exposed to across the lifespan change the way older and younger speakers of English articulate signals in verbs and nouns.

Why verbs and nouns? First, verbs and nouns, by token count, tend to be equally frequent in English. At the same time, nouns are more productive than verbs: there

[1] in relation to all other words
[2] in relation to all other words

are more different noun types than verb types, and the difference increases as the sample size grows. Because the infrequent noun types are more likely to occur in specific contexts only, they are distributed irregularly across speech samples and contexts. Consequently, all speakers encounter and learn more new nouns than new verbs throughout adulthood, and speakers' individual experiences with infrequent nouns can vary enormously.

Further, as discussed in section 3.1.1, the argument structures that support English verbs are less variable than the argument structures in which nouns occur. Verbs are subcategorized by fixed grammatical conventions and are not arbitrarily interchangeable in context (Levin, 1993). Nouns, by contrast, can occur within any given argument structure where nouns are expected. However unconventional or nonsensical a message may come out of placing a random noun in a noun argument frame, the sentence will not be ungrammatical. Due to this, the uncertainty about the lexical contexts in which verbs occur is far more fixed by convention, so that the uncertainty involved in verb production ought to, on average, be lower than the uncertainty involved in noun production.

Finally, the gradual changes in the distribution of information that occur across adulthood ought to affect nouns and verbs differently. If signals are adapted in response to the uncertainty associated with the context, changes in the way older speakers produce nouns and verbs ought to reflect these differences. The results of the analyses presented in this chapter indicate that they do.

> **Main findings:** head and tail growth in 'Zipf-like' distributions seem to follow from learning; signal production seems to reflect dynamic redistribution of information across words and the contexts they appear in; patterns of articulation in word-initial and word-final positions appear to reflect the functional load on verb and noun morphology

6.1 Is speech development a lifelong process?

Does adult speakers' experience shape the way speakers articulate signals in words? Speaking obviously involves learning: children learn to speak from the input they are exposed to in meaningful interactions with their caregivers. Learning allows them to gradually master increasingly fine-grained patterns of temporal and spectral variation that form the speech conventions shared by their community. Because language development is a gradual process, 'developing' language is accompanied by a great deal of variability in the way utterances, words, and sounds are produced over time and in individuals. While the rapid progress in developing speakers makes

it evident that form is a matter of learning and learning is a matter of exposure, variability of form in spontaneous adult articulations is rarely explored from this perspective. Is speech development a lifelong process, or does it settle at some point because there is nothing left to learn?

6.1.1 What do speakers learn when they learn to speak?

It is a truism that as talking involves effortless command over sounds and words and grammatical constructions, learning to talk must entail learning to recognize and articulate syllables and sounds and from those words and grammatical constructions. The building-block metaphor, however, sits ill with two critical properties of human development. First, learning is an organizing process that evolves under developmental and experiential constraints. And second, a critical feature of learning is that it leads to the development of preference. The development of preference is important in complex environments because it leads to choice (and uncertainty) reduction, which in turn enables humans to notice fine-grained differences in our environments and learn to make more specific distinctions and develop more specific preferences (Kuhl et al., 2006). Similarly, human communication appears first possible because humans **learn not to attend** to large parts of the input available in our environments and instead develop the ability to selectively attend to context-relevant sources of input (Rivera-Gaxiola et al., 2005; Best and McRoberts, 2003; Tsao et al., 2006; Werker and Tees, 1984).

Why are preference and choice reduction important in language learning? While the infant human can theoretically learn from any signal that is sufficiently structured and available to its senses[3], the extreme variability of vocal signals it is exposed to does not provide the kind of structure that would allow learners to maintain their expectations over time (see section 5). In other words, a mind that has not developed the ability to filter out those parts of signals that are unstructured and thus irrelevant to the user allows noisy signals. To contain the chaos of sensory experience it faces, the human infant must first establish a set of expectations that are consistently reinforced by patterns of vocalizing (and other) behavior it is exposed to (Kuhl, 2004; Kuhl et al., 2006; Mattys et al., 1999; Werker and Tees, 1984). In addition to this, early learning from noisy input seems to benefit from developmental

[3]The human infant is born with a minimal neurophysiological specification, its brain at the peak of synaptogenesis. **Synaptogenesis** describes the process by which connections in the brain neurons are formed. The process is initialized by the *exuberant synaptogenesis*, a growth burst that occurs in early development. Later development then aims to reduce connectivity through competition (pruning). Connectivity correlated with processes that are not used during this period of pruning will likely not develop later on in life. The brain user, at the onset of this process, has not developed much preference, and the input she is exposed to is largely unfiltered. The upside of this excessive wiring is that it allows for maximal adaptation; a newborn baby can theoretically learn 'almost' anything. The downside of super-connectivity is its metabolic cost and its lack of efficiency

constraints: perceptual and memory limitations may allow young learners to discover and attend to aspects and dimensions of sensory input that are not easily learned later in life. This idea was introduced as the 'less-is-more' hypothesis (Newport, 1988; Elman, 1993)[4], and the principle seems to hold across sensory modalities. For instance, computational models of early vision development suggest that infants' underdeveloped vision benefits the later development of sensitivity to binocular disparities that are useful for motion detection (Dominguez and Jacobs, 2003).

In language learning, developing speakers appear to benefit from the protracted development of the prefrontal cortex that is typical to humans (Fuster, 2002; Ramscar and Gitcho, 2007; Thompson-Schill et al., 2009; Teffer and Semendeferi, 2012). This seems to allow children to extract regularities from speech that are not available to adults who have learned to preferentially attend to individual words in sequences (Arnon and Ramscar, 2012; Ramscar et al., 2013b; Ramscar et al., 2013a; Ramscar et al., 2013c; Finn et al., 2014; Hartshorne et al., 2018). More recent analyses suggest that the reduced signal quality during prenatal exposure allows learners to extract signal dimensions that help them structure the noisy time-variant signals in speech. Exposure to low-pass filtered sound in the amniotic environment appears to benefit learning. Computational simulations indicate that exposure to severely degraded auditory input can induce neural development by allowing learners to extract signal dimensions that, among other, help them develop stable temporal expectations (Vogelsang et al., 2022). In particular, degraded input seems to help speakers extract signal dimensions that develop at slower timescales (are set further apart than words and phones), such as prosodic phase markers. For example, lengthening of phrase-final vowels, changes in fundamental frequency, and pauses, all of which tend to be periodic and all of which infants use in early segmentation (Bard and Anderson, 1983; Fernald, 1989; Männel and Friederici, 2009; Männel and Friederici, 2011; Sundara and Scutellaro, 2011; Männel et al., 2013; Skoruppa et al., 2013; Sundara et al., 2015). As discussed in the previous chapter, periodicity is a necessary precondition to establishing consistent expectations, allowing speakers to structure signals and extract more fine-grained information.

What do preference development and choice reduction have to do with learning in general and learning to speak in particular? Learning is a process that seeks to minimize the difference between learned expectations (the model of the environment) and the experienced environment. Any perception/experience, and by implication, any learning that can follow from it, is facilitated by the extent to which extant expectations conform with current and future exposure. Because new expectations are acquired in the context of acquired expectations (i.e., learning is discriminative), for learners to learn about the details of their environments, some aspects of this

[4]Note that considerations on the 'early advantage' and the details of its implementation proposed by (Newport, 1988) and (Elman, 1993) are not undisputed (see e.g., Rohde and Plaut, 1999).

environment must be sufficiently predictable to provide context to new learning. Accordingly, exposure and adaptation to **consistent** sources of signal (behavior) in the immediate linguistic environment gradually alter learners' expectations and lead to specialization. Specialization is characterized by unlearning global patterns of crying, cooing, and babbling at the benefit of learning more specific patterns of articulation that confound local speech habits (Kuhl, 2004; Kuhl et al., 2006; Mattys et al., 1999; Werker and Tees, 1984).

How does this relate to what the adult learns from exposure? This characterization of learning conflicts with the idea that speakers construct the signal from an alphabet of articulatory gestures. Instead, learners appear to set up a common set of expectations about the structure of signals they are exposed to (i.e., they build a model of the spoken environment). This model allows them to make sense of the increasingly complex and fine-grained patterns of variation they extract from speech signals. One fundamental structural expectation speakers establish seems to be related to the temporal and prosodic structure of natural speech sequences. The prosodic structure is related to several signal dimensions speakers learn to extract during early infancy (pause, variations in fundamental frequency, and phrase final lengthening). Speakers' ability to identify and discriminate segmental and gradient acoustic contrast seems to be bounded by these initial expectations. For example, speakers appear to be less sensitive to the intelligibility of signals that occur at the initial boundaries of prosodic phrases. Quality degradation in the initial segment of prosodic phrases mostly goes unnoticed and does not seem to affect phrase intelligibility (Doelling et al., 2014). Also, speakers' ability to extract short words and phonemes from the signal seems to depend on the speech rate of the preceding phrase, and short words and unstressed segments are easily overheard when the context rate is manipulated (Dilley and Pitt, 2010; Baese-Berk et al., 2019). Temporal bias can alter speakers' interpretations of ambiguous syllable structures (e.g., whether a phrase is identified as *tie murder bee* or *timer derby*) (Morrill et al., 2014), and affect recall (Lamekina and Meyer, 2022). Interestingly, the extent to which the phonotactic structure aligns with speakers' expectations, in turn, seems to affect the rates at which speakers segment the upcoming part of the signal and their ability to extract syllables and morphemes in the course of the conversation (Finn and Hudson Kam, 2015). This **uncertainty effect** can also be induced artificially by disruption of the relevant brain circuits (Smalle et al., 2017). The cumulative effect of such misalignment in rate and expectation (and expectation and rate) appears to lead to a dynamic reorganization in the perception of acoustic-phonetic detail. Speakers' sensitivity to speech sounds and boundaries changes systematically across phonetic categories to accommodate the uncertainties related to the signal source (e.g., non-native speech) (Xie et al., 2017; Xie and Myers, 2017).

All of the above suggests that on one hand **phoneme or morpheme informativeness** depends on which parts of the acoustic signals are segmented and extracted by listeners, and **can vary considerably with the length and the temporal structure of the phrase**. On the other hand, the idea that much of what is articulated never gets segmented and that speakers only attend to acoustic-phonetic variation that is consistently discriminated in the context of the rate suggests that **learning to ignore certain features of the input is an important aspect of communicative development in both adults and children**. The difference between children and adults seems to be that the adults' temporal expectations seem to be relatively fixed and operate on shorter timescales (i.e., adults' perception can re-scale dynamically at syllable and phone boundaries). In comparison, children seem to rely more on prosodic cues at the utterance boundaries that unfold on slower timescales. In the previous chapter, we presented analyses that suggest that temporal expectations will change continuously across adulthood. The analyses from chapter 5 suggest that the temporal resolution in utterance-final positions becomes increasingly fine-grained as speaker experience grows. This raises a question: do such changes manifest in the way older speakers articulate acoustic-phonetic signals in words? How can we measure this?

6.1.2 Are speech form dynamics a functional response to uncertainty?

In previous chapters we mentioned that changes in articulation rates in English can be linked to vocabulary development across the lifespan. The development of increasingly diversified vocabularies is a by-product of learning and specialization. **How will lifespan learning interact with the speech signal structure?** As summarized above, it appears that systematic choice reduction (i.e., bias) that follows from **preference** (acquired patterns of generalization) is the crucial condition for communicative specialization and development unique to human kind. Communicating an increasingly diverse repertoire of context-specific knowledge depends on predictability. **Predictable forms** allow speakers to efficiently and reliably transmit the messages related to the increasingly specific distinctions context-specific knowledge allows one to make. The findings summarized in chapter 3 suggest that predictable forms are maintained through convergence in the shared patterns of usage. We explained how regular patterns of word distributions discriminate between communicative contexts, which in turn allow speakers to interpret and learn rare and novel words in context (Ramscar, 2020; Ramscar, 2019). We stressed that social evolution and the evolution of human communicative codes involve cooperation and the maintenance of mutual predictability set by shared (and shareable) systems of convention (cf. Peters and Adamou, 2022). Speakers' ability to remember and efficiently transmit a rapidly growing number of increasingly specific distinctions/vari-

ants is relative to the predictability of the communicative contexts that sustain them. Accordingly, speaker communities maintain a set of abstract expectations (prosodic patterns, syntactic and semantic conventions, periodic vocal signals of some kind) that do not change substantially over a human lifespan. These stable and shared communicative conventions can be infinitely extended/varied to accommodate for the accumulating differences between the specific communicative demands of individual speakers at different levels of experience. In other words, **the productivity of human communicative codes appears to hinge on the predictability of a common functional model and the reduction of uncertainty that follows from it**. How does this manifest in speech?

6.2 From signals to forms: The information structure of spontaneous speech

Speech patterns we learn to know as phrases, words, and sounds, are form variants discriminated by the rates at which quasi-periodic sequences of air puffs modulated by articulator configurations occur in speech signals. The rates are correlated to the three main amplitude modulation frequencies at which boundaries of prosodic phrases, syllables, and phonemes unfold. The rates at which the boundaries of acoustic segments occur in physical signals seem to be correlated to the rates at which acoustic events in speech signals occur in synchrony with rhythmic or repetitive patterns of neural activity in the human brain (neural oscillations) (Luo and Poeppel, 2007; Giraud and Poeppel, 2012; Meyer, 2018; Poeppel and Assaneo, 2020).

The processes that connect symbols, signals available to senses, and rhythms of the mind are stubbornly elusive to the language and brain scientist. Virginia Woolf, in a letter written to Vita Sackville-West in 1936, lucidly identifies it as rhythm (Woolf and Nicolson, 1975, p.)

> As for the *mot juste*, you are quite wrong. Style is a very simple matter: it is all rhythm. Once you get that, you cant use the wrong words.... Now this is very profound, what rhythm is, and goes far deeper than words. A sight, an emotion, creates this wave in the mind, long before it makes words to fit it; and in writing (such is my present belief) one has to recapture this, and set this working (which has nothing apparently to do with words) and then, as it breaks and tumbles in the mind, it makes words to fit it.

How do we meaningfully relate the ways in which waves in the mind turn variation in continuous signals to discrete representations? On one hand, we have a corpus of words that in their transcribed form, approximate words spoken in isolation. On the other, words seem to achieve their purpose in sequences. The sequences are said to be governed by generative processes, but in actual speech seem to approximate a system optimized for discrimination. Words take form in the context of prior experience, and this prior appears to be temporal and discriminative. How can this be expressed in a model?

As we noted above, talking involves adaptation and management of expectations. A misalignment between speakers' explicit expectations and the implicit information they extract from the signal leads naturally to the notion of information processing; i.e., what is expected and how are these expectations organized? Human information processing is a feat of discrimination (as is all information processing (cf. Shannon, 1948)). By implication, speaking ought to entail learning to detect and initiate meaningful changes in the signal that discriminate between context-relevant alternatives or any meanings associated with them - speech contrasts. This raises three obvious questions: (1) What defines a speech contrast? (2) What makes it context-relevant? And finally, and perhaps most importantly, (3) What is **context**?

First off, to define contrast in terms of discriminative information, consider the following vocal interaction:

1. A: SILENCE
2. B: *aaaaaaaaaaaaah*
3. A: SILENCE

Since (2) is clearly not confusable with (1), we can say that the information provided by the signal is set by the absence of spectral variation in (1), so that any change that interrupts the silence is sufficient to identify (2) as a speech fragment of type **not a silence, a sound**. In a world that only had one alternative to offer the state of one would always provide enough information about the other, there would be nothing left to say. Humans, however, have a lot to say, and the number of things we find remarkable enough to mention increases rapidly over time. Accordingly, a critical point of a discriminative approach to human communication is that learning from exposure yields more informative variation because more is different in some informative (meaningful) way to someone.

By consequence, silence in the speech signal provides context to a category of vocal gestures of type *sound* where the discriminative contrast is set by the change

in the relationship between the spectral maxima (*aaah* and *oooh*) and durations (*aaaaaaaaaaaaah* and *aaaah*) – **speech contrast is the dimension that discriminates a specific sound from other possible competing sounds**[5]. Phonemes, and other speech form abstractions, by contrast, capture the common ability to articulate *eeeeeh* when \i\ is meant. The abstract representations (e.g., phonemes) are contrasts embedded in a minimal unambiguous auditory context to preserve the phonemes' discriminability when they are produced in isolation[6]. In real-life communication, however, the speech contrast variety that the isolated form aims to approximate is seldom encountered in isolation. Instead, speech contrasts appear as parts of structured sequences, such that the first **contrast set by the absence** of sound distinguishes two categories: periods of sound (articulations) and periods of silence (pauses) that unfold in time.

1. A: *aaaaaaaaaaaaah* SILENCE
2. B: *oh* SILENCE *oh* SILENCE *whohohoho* SILENCE

The sequence makes allowance for the constraint set by the average/human communicative bandwidth. Because the number of possible messages is theoretically unbounded and the amount of information (discriminative contrast) that the human brain can process simultaneously is limited, expectations need to be managed incrementally[7]. This adds a layer of complexity to the speech problem: How do speakers know whether and what [not] to expect at a given increment?

The findings summarized in previous chapters suggest that the answer lies in the structure of *shared codes*: human languages are structured by systematic patterns of co-variation between words, morphemes, and phonemes. This structure seems to maintain a memoryless distribution of nested distributions of words, morphemes, and phonemes. The distribution allows speakers to build sufficiently similar expectations models independent of the differences between the individual samples they are exposed to and learn from. These distributional regularities discriminate between classes of semantic and morpho-syntactic 'events'. Events from these classes differ in the degree to which they maintain the predictability of speech sequences by providing context and the degree to which they distinguish between individual messages. The distributions appear to adapt to the demands of the diversifying

[5] sounds one can reasonably expect to interrupt silence in conversations

[6] What we remember is everything that we ought not to expect from the alternatives. This must not necessarily be a change in the auditory signal. Auditory-visual illusions show that people merge information across senses when interpreting speech sounds. For example, hearing the sound 'ba' while seeing a mouth articulate 'ga' often leads adults to interpret the sound as 'da,' a blend of the two. What we recall having heard includes the uncertainty over the unspecific variation in silence after the alternatives have been eliminated - the model form.

[7] such that not all contrasts are expected at each increment

environments through the transmission process itself[8]. The transmission process results in distributed word clusters that sustain sets of lexical alternatives that differ in some audible dimension and guarantee minimal differences between samples individual speakers are exposed to. In other words, the system structure seems to balance out the competing requirements of collective convergence (i.e., mutual predictability) and individual divergence (i.e., discriminability). It guarantees that all speakers are exposed to sufficiently similar distributions of sufficiently distinct speech form variants at each increment. The nested structure implies a degree of modularity and independence between the lexical subsets: speech forms can diversify in context without breaking the structure. The modularity suggests that global functional pressures, such as for example population growth, demographic changes or technological development (see e.g, Klingenstein et al., 2014; Iliev and Axelrod, 2016; Soni et al., 2021), can be mediated by local constraints (by further subcategorization and an increase in local contrast, i.e. burstiness). We will examine this idea in the context of lifespan speech production: do permanent changes in patterns of articulation reflect the extent to which local constraints (set by grammars and other speaking habits) mediate global pressures (prompted by learning from increasingly diversified signals)?

Speech corpus analyses provide evidence to support the idea of local, context-specific adaptation, showing that across languages, both word morphology and subtle variation in the way words are articulated corresponds to functional pressures in context (Wedel, 2012; Blevins et al., 2016). Functional approaches define the communicative function in terms of *informativeness*. Informativeness, often formalized in terms of the information-theoretical quantity *information content*[9], approximates the extent to which articulated signals reduce the uncertainty that has not already been resolved by the context in which the articulated signals occur. Speakers appear to consistently enhance signal dimensions that discriminate between phonetically similar words that appear in similar contexts, and attenuate dimensions that do not contribute to discrimination (Wedel et al., 2013b; Wedel et al., 2018). Consistent with

[8] whether unique forms are remembered and transmitted depends on the demands of communicative contexts in which they are used

[9] Information content of a word w from a set of words $W = word_1, word_2, word_3, ..., word_n$ discriminated by a context which can be loosely defined as a finite class of words as in $-log_2 P(word_i|word_1, ...word_n)$, co-defined as final part of an utterance in context as in $-log_2 P(word_i|word_1, ...word_{i-1}, CONTEXT)$, or left unspecified as in $-log_2 P(word_i)$. The context is usually defined as the probability of a word occurring within an n-gram (e.g. Piantadosi et al., 2011), a document (e.g. Wedel et al., 2013b), a syntactic structure (e.g. Levy, 2008), or a class of words not distinguished by acoustic or orthographic features (e.g. Priva, 2008) (with latter approach raising the question whether an absence of difference can be taken as evidence for similarity in discriminative codes). As we have stressed at various points in the introductory chapters, current implementations of information content disregard the fact that the probability space of W (or $CONTEXT$) co-varies with the sequence position and speakers' experience, and that one of the major challenges for communication and the implementations of information-theoretic measures involves accounting for the probabilities of unattested events at different points in time (cf. DeDeo, 2018).

this, we show in chapter 4 that deviant phonemes lead to a quantifiable improvement in the distribution of discriminative contrast at the word boundaries: word-initial segment variants yield probability distributions that approach the theoretical limit of communicative efficiency (Shannon, 1948). Remarkably, words from syntactic categories that differ largely in type-token-ratio and thus the distribution of word-initial contrast presupposed by the model (the dictionary form), converge on nearly identical distributions of word initial phonemes. We see similar patterns in the distribution of initial segment duration. All of the above suggests that apparently incidental form variation in spontaneous speech is a structured, productive aspect of human communicative behavior that increases the discriminability of speech signals.

All of the above suggests that the speech signal is far more complex in its structure than text-biased intuitions would initially suggest. It also implies that lifelong exposure will make it increasingly complex and structured. Why? As we have emphasized here, learning is essentially a discriminative process that leads to changes in the way the environment is experienced - at lexical, sublexical, or any given level of description. The signal structure that follows from discriminative learning guarantees that learning is a lifelong process (Ramscar et al., 2014; Ramscar et al., 2017). Taken together, these considerations allow us to make a number of concrete predictions about communicative development across the lifespan. They suggest that not only does lexical knowledge increase across the lifespan but also that all aspects of linguistic behavior, including speech perception and production, should be subject to the same systematic patterns of change over time. In what follows, we address this in particular.

6.3 Lifespan development: Do articulations really get slow and hazy as people get older?

Earlier, we asked what speakers learn when they learn to speak. While answers to this question vary, the usual consensus appears to be that the process ends at some point in adolescence. Given that we know that language models change over time and that perception and production will be conditioned on these models, it seems unlikely that articulation would fossilize at such an early age. In line with this, results from studies that target the effects of speaker age on vowel production, speech rate, and pause duration capture systematic changes in aging speakers' articulations. When changes in temporal and segmental variation are considered together, the effects at first appear a bit opaque and inconclusive. Lifetime changes in speech rate follow an inverse u-shaped pattern, where speech rate increases with speaker age, peaks in middle-aged speakers, and from there on decreases, on average, becoming increasingly variable across speakers and contexts (Jacewicz

et al., 2010; Hazan and Pettinato, 2014; Bona, 2014). Notably, these differences seem to not replicate entirely across languages: Gerstenberg et al., 2018 report increasing articulation rates in older speakers of French and decreasing articulation rates in older speakers of German. Also, experienced public speakers' speech rates seem to increase across the lifespan (Quené, 2005; Hunter et al., 2012). Simultaneously, the average pause duration remains stable, showing only small local decreases across early and late adulthood (Redford, 2013; Neuberger, 2013; Bóna, 2011; Gerstenberg et al., 2018). While older speakers pause more frequently, the duration of individual pauses and with it the variability in average pause duration across individual speakers decreases over time (Demol et al., 2006; Hazan and Pettinato, 2014; Bona, 2014).

By contrast, the individual variability in segment duration and vowel space resolution appears to increase. While there is a general tendency for vowels to get longer and less concentrated through a drift towards the vowel space periphery (Fletcher et al., 2015), longitudinal data suggests that vowel space expansion becomes increasingly independent from vowel duration across adulthood (Gahl and Baayen, 2019). Notably, Watson and Munson, 2007 find no age effect on the size of the vowel space in words spoken in isolation, which indicates that age-related differences in segmental variation pertain to connected speech and may be prompted by the context in which words appear. In sum, changes in the temporal resolution and the acoustic form of signals articulated in context and signals articulated in isolation appear to follow distinct patterns. How does context affect the increasing misalignment in how speaker cohorts realize speech forms in time?

One possible explanation for the inconsistencies in older speakers' production rates could be that the effects reported in the literature are a result of pooling over speech contrasts and contexts that fulfill distinct functional requirements. We mentioned earlier that experience leads to different patterns of development across signal dimensions. Some signal dimensions converge over time and individual speakers, while others diverge. In regularly distributed, periodic aspects of the signal, such as pauses, vowels, function words, and regular patterns of inflection (e.g., tense and number markers), learning is asymptotic; after the speaker has observed a sample of a reasonable size, there will be nothing new left to learn. Signal dimensions that are irregularly distributed in time, by contrast, become increasingly discriminated (i.e., learning leads to 'phase' transitions). The rate at which people detect noticeable changes in signal increases, and the extent to which fine-grained differences in the signal dimensions are perceived change over time (cf. Tuller and Kelso, 2018). What exactly leads to phase transitions in learning (i.e., when and why people transition from one temporal reference scale to another reference scale) is not entirely clear. Following analyses presented in chapter 5, we have suggested that re-scaling in speech is sensitive to the amount of variable error (the variability of estimates of

some objective magnitude or event probability measured by their average deviation) in periodic aspects in signals. For the purposes of the analyses presented in this chapter, we shall focus on vowels, silent and filled pauses and contrast the realization of word-final and word-initial boundaries, which seem to, on average, fulfill different functional purposes in communication (Ramscar et al., 2013e; Ramscar et al., 2018; Wedel et al., 2018; King and Wedel, 2020).

On one hand, this suggests that pauses, vowels, function words, regular suffixes, and markers of prosodic phrase boundaries can all, to varying degrees, **contribute to the management of expectations (and uncertainty)**. It implies that recurrence rates and the magnitude of noticeable changes in all of the above will co-vary systematically with the uncertainty about the signals that follow and **determine the degree to which articulations contribute to discrimination between competing signals** (i.e., whether articulations are perceived as signal or discarded as noise by the individuals brain). Consistent with this, vowel and silent interval[10] durations are shown to serve as cues to the uncertainty of the upcoming part of the signal: longer vowel durations facilitate the prediction of less frequent word forms (Kemps et al., 2005; Salverda et al., 2003; Shatzman and McQueen, 2006) and lexical arguments (Jurafsky et al., 2001; Bell et al., 2003; Andruski et al., 1994). Similarly, experiments show that speech rate inconstancy, pauses, and disfluencies make upcoming speech more fluent and intelligible (Bosker et al., 2013), the messages easier to remember (Fraundorf and Watson, 2011; MacGregor et al., 2010; Corley et al., 2007; Diachek and Brown-Schmidt, 2022) and lead speakers to choose unfamiliar and less predictable options (Arnold et al., 2003; Arnold et al., 2007). Meanwhile, hyperarticulation of segmental contrast, which typically coincides with faster speech rates, increases the discriminability of lexical alternatives in context (Wedel et al., 2013b; Wedel et al., 2018). That is, regularly distributed aspects of signals seem to regulate expectations, while irregularly distributed aspects of signals help discriminate between competing alternatives in context. The results presented in the foregoing chapters suggest that the rate at which word sequences resolve uncertainty remains relatively stable as speakers' experience increases. Simultaneously, the diversity of lexical samples they are exposed to increases continuously throughout adulthood.

Where exactly do we expect differential impact of speaker experience on articulation? Vowel space, speech rate, and voice analyses make up a large part of research contributions on changes in articulation in healthy aging (for review see Tucker et al., 2021). These investigations usually focus on differences in a particular sound dimension (and often a particular speaker), disregarding changes in the temporal resolution of signals. The focus on differences between variants

[10] pause and voice onset time

raises the question of which aspects of the signal the analyst expects not to vary. As we have stressed in previous chapters, it appears that a meaningful definition of *variances* is only possible once we have established a clear definition of *invariance* (cf. Bürki, 2018; Dienes, 2008). The challenges of measuring 'meaningful' variation in a particular signal dimension is well exemplified in changes in formants, which distinguish between vowels and reflect how different articulator configurations affect the vocal tract resonance in different frequencies (F1, F2, and F3[11]). Formant measurements are usually extracted from the vowel mid-point. Vowel durations, however, vary consistently with the following consonant: vowels followed by voiced consonants are significantly longer than vowels followed by voiceless consonants. These differences, however, are often smudged by carry-over effects and seem to not be perceived particularly accurately by speakers (Fowler, 1992). We have argued in the foregoing that the extent to which changes in frequencies are registered by speakers depends on the rate at which the signal is segmented and that rates and informativeness of changes seem to vary with the utterance position and speaker experience. All of this seems to suggest that the signal the phonologist is measuring may be very different from the signal the speaker is producing and, again, different from the signal the listener is extracting. How can we control for contextual variability when measuring lifespan changes in production? Where do we expect (or not expect) changes related to adult experience to occur, and how do we expect these changes to instantiate?

Is there a difference between experience-related and pathological change in older speakers' performance? Older speakers typically take longer to recall words in lexical retrieval experiments. The differences in the extent to which older speakers experience more lexical uncertainty in recall tasks compared to younger speakers appear to vary with the lexical category (i.e., proper noun recall elicits longer response times, while common noun recall does not) and the amount of support provided by the context. These differences tend to disappear and are sometimes reversed when words are presented in context (Sommers and Danielson, 1999). Older speakers' performance generally tends to be interpreted as evidence of a decrease in cognitive fitness. That is, even when older speakers fail to produce longer response times in context, their performance is interpreted as an 'inferior' strategy, reflecting an over-reliance on the syntactic structure that serves to counter the effects of cognitive failure. Similarly, while faster speech rates in younger speakers' tend to be interpreted as evidence of healthy cognition, increasing speech rates in older speakers are explained away as an attempt to catch up with information lags caused by respiration rates[12] (Gerstenberg et al., 2018). Notably, the onset

[11] formant frequency bands capture the average number of cycles/changes in amplitude a sound wave completes in a fixed time interval of 1 second; F1 centers at $\sim 500Hz$ approximately, F2 at ~ 1500, and F3 at ~ 2500

[12] Gerstenberg et al., 2018 note that *some older speakers may inhale more often, but may compensate for the reduced respiratory capacities by an increase in articulation rate to maintain information*

of the decline in performance is measurable in speakers in their twenties and progresses from there on (Ramscar et al., 2014).

Changes in communicative behaviors, however, must not necessarily reflect senility. While the likelihood of physiological illness and social isolation are definitely higher in older individuals than in younger individuals, there are, in general, more healthy older adults than sick older adults in modern societies[13]. World-wide prevalence of dementias in 65- to 69-year-olds is estimated at around 1%, increasing exponentially over the later decades, such that up to 30% of over 90-year-olds suffer from some form of dementia. This seems to suggest that approximately 99% of people aged 65-69 do not suffer from a measurable, pathological decline in cognitive performance, and approximately 70% of all people older than 90 years are unaffected by dementia. Healthy aging does, however, seem to affect the way people respond to information. This is not entirely surprising, as changes in behavior often tend to reflect changes in the information structure of the experienced environment. Consistent with this, meta-analyses and empirical studies support the idea that the differences in the younger and older speakers' performance reflect changes in the structure of the cumulative samples they are exposed to (Ramscar et al., 2014; Ramscar et al., 2017), and changes in speaker performance pertain to signal dimensions rendered uninformative by experience and local changes in the distribution of uncertainty. For example, the turnover rates in proper nouns relative frequencies (e.g., brand names and persons names) tend to be much higher both across regions and decades (see also chapter **??**), so that speakers' estimates of the reliability of arguments appearing in name-frames will change as speakers get older (and experience more decades, and often more regions). In other words, individuals' certainty about their own models and the global distribution of uncertainty ought to increase with experience, while at the same time speakers' certainty about other peoples models and the various local distributions of uncertainty ought to decrease.

In line with this, older speakers are reported to experience more partial retrieval blocks/tip-of-tongue states in proper nouns than in common nouns (Evrard, 2002), while Seifart et al., 2018 find more disfluencies preceding nouns compared to verbs across languages. The analyses presented in chapter 4 suggest that the likelihood

density and discuss decreasing pause durations in older speakers as *a way to save energy and to compensate for the greater effort that speaking demands with increasing age*

[13] There is also more variance in individual living conditions, and this individual variance may increase along with older speakers' performance. The likelihood of illness and social isolation is higher in those parts of populations where the likelihood of poor healthcare and social neglect are high. Simultaneously, the likelihood of having acquired a relatively high standard of living and more certainty about most basic existential problems, in general, is higher in older than in younger cohorts. In other words, differences between individuals ought to be exacerbated in the older cohorts in modern-day societies (which are characterized by an unequal distribution of wealth and other, often related, resources).

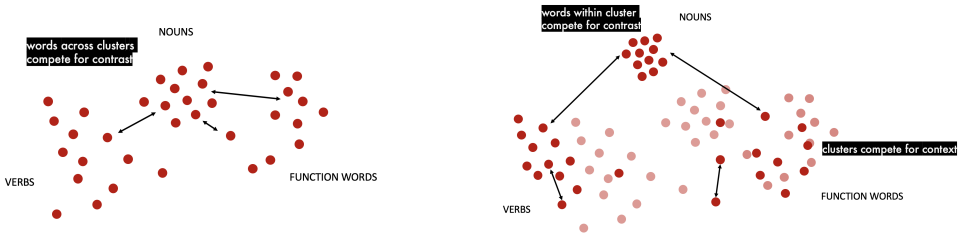

Fig. 6.1: Schematic illustration of the predicted dynamics of the distributional structure in time: Grammatical constraints will impose limits on the variability of lexical contexts. This, in turn, will mediate the competition between words from different semantic categories. Learning will increase the distance and decrease lexical competition between clusters of words that do not share contexts (e.g., proper nouns and gerunds). Words that are well discriminated from other words by the context compete for word-initial contrast with the cluster cohort (e.g., proper nouns with proper nouns they share contexts with). The acoustic features of words that provide context to words from multiple grammatical and lexical classes (e.g., function words and other contextually dispersed words) will become increasingly uninformative and increasingly variable in duration as the experience increases. Word forms that lie between the two extremes (of contextual dispersion) compete for both context and contrast with other word forms, which increases the functional load on acoustic contrasts at both word-initial boundaries (that are more likely to increase the differences between lexical forms (King and Wedel, 2020)) and word-final boundaries (regular suffixes that subserve the discrimination of morpho-syntactic categories, i.e., contexts (Blevins et al., 2016; Blevins et al., 2017; Ramscar et al., 2013e)).

of form variation increases with the diversity of lexical contexts words appear in. In line with this, Tremblay et al. (2008) show that articulation trajectories appear to be highly context-specific: articulations do not generalize across different utterances even when the utterances are matched for kinematics, indicating that articulations are not modeled on a finite set of articulatory gestures. This implies that variation in articulated signals may reflect the degree of local habituation – where more variation implies more uncertainty and less habitual trajectories, resulting in more diverse patterns of articulation (Tomaschek et al., 2018). This raises the question of whether articulations vary systematically with context. In earlier chapters, we introduced a definition of context that discriminates between nested clusters of grammatical, lexical, and sublexical contrasts by learning from patterns of distribution. This distributional notion of context implies that some clusters will become increasingly diversified within themselves over time by increasing the lexical competition within the cluster. Meanwhile, other clusters of words from lexically less productive categories will become increasingly discriminated by the contexts they appear in. Figure 6.1 illustrates this progression on the example of word form competition between words from different lexical clusters.

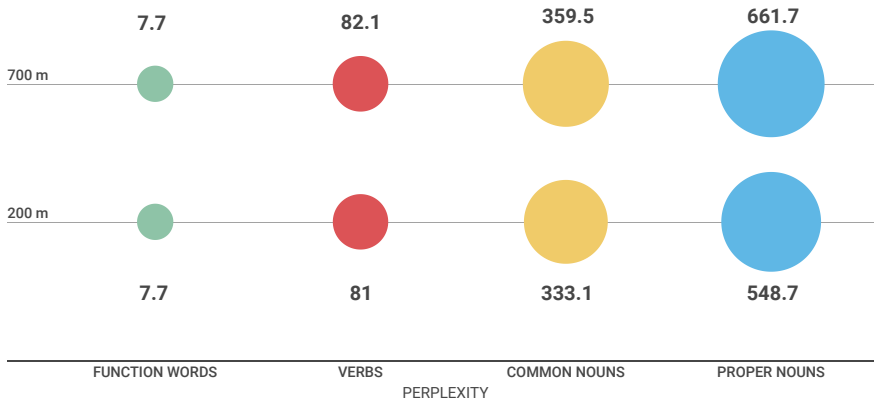

Fig. 6.2: Top row: Perplexity of function words, verbs, common and proper nouns from 200 million words (bottom line), and 700 million words from the Google Books trigram corpus (top line). A substantial increase in sample size only leads to minor increases in lexical perplexity in verbs (81 to 82.1) compared to common (333.1 to 359.5) and proper nouns (548.7 to 661.7) (Ramscar et al., 2014). Given that English grammars, with their relatively fixed word order, impose constraints on the degree to which words from different categories can reorganize across sequences (i.e., the combinatorics are relatively fixed), we expect conditional probabilities between nouns and the contexts they occur in to increase the differences between individual speakers' models, while models of transitional probabilities between verbs and their precedent arguments ought to be more stable in time.

This theoretical model of 'speech development' suggests that variations in the way speakers realize speech forms in context reflect the distinct rates at which experience increases the variability (and uncertainty) across functionally distinct distributions of linguistic cues. For example, when word categories are considered independently, the perplexity[14] of different word categories (e.g., verbs and nouns, see Figure 6.2) does not increase in synchrony (at least according to corpus counts). Accordingly, the regular distribution of parts-of-speech categories implies systematic differences in the co-variate structure of lexical samples containing words from these categories[15]. As noted above, previous findings have shown that response differences between age cohorts are well explained by the systematic changes of co-variation patterns in the samples speakers are exposed to and learn from across the lifespan (Ramscar et al., 2014; Ramscar et al., 2017). The irregular distribution of words from the productive lexical categories (open classes) guarantees that

[14]The perplexity PP of a discrete probability distribution p is defined as $PP(p) = 2^{H(p)} = 2^{-\sum_x p(x)\log_2 p(x)}$

[15]i.e., English grammars constrain the word order, so that the rate at which part-of-speech classes and the lexical types they host occur mutually is likely predicted by the rate at which they occur individually

vocabularies will expand dynamically across adulthood. Given the foregoing, we expect the distribution of functional load across word and utterance positions to develop differently in verbs and nouns. If phonetic variation reflects uncertainty, the increasing misalignment in the distribution of information across words and the lexical contexts they appear in should affect how younger and older speakers articulate verbs and nouns.

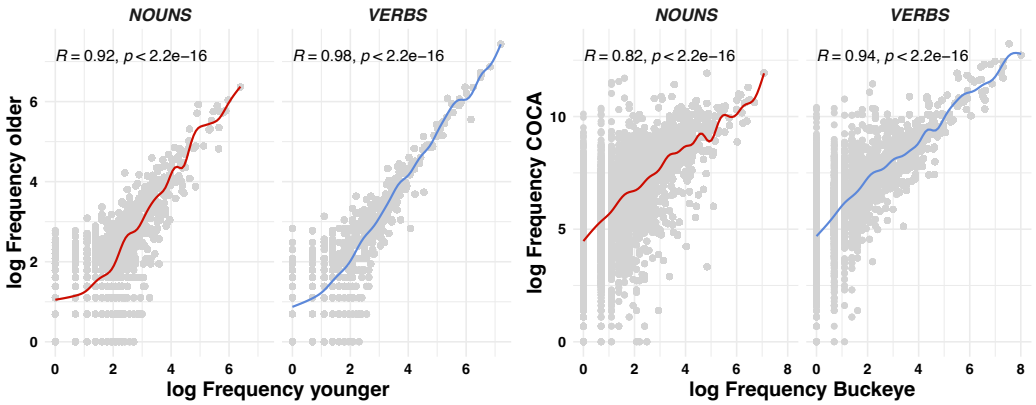

Fig. 6.3: Bottom: Correlations between log frequencies of nouns and verbs produced by older and younger speakers (left), and noun and verb frequencies from the Buckeye corpus (300 000 tokens) and the much larger spoken part of the Corpus of Contemporary American English (80 m tokens) on the right. Older speakers produce low-frequency words more often. Note that the 'misalignment' between cohorts/samples is not limited to the tail of the distribution. There are large differences in the relative probability of words from all frequency registers. The larger corpus overestimates the local probability of low-frequency words and underestimates the register-specific re-ranking of higher-frequency words. As we have noted in chapter 2, re-ranking within sub-categories can serve as a systematic (i.e., predictable) source of error and thus facilitate learning and adaptation. In 'closed' categories (categories stratified by orthography and grammar), this re-ranking in aggregate distributions will likely lead to Pareto- or Yule-like shapes (increase contextual diversity, (cf. Klingenstein et al., 2014)). In more productive word categories, the re-ranking will likely increase the utterance length.

In the following section of this chapter we examine how these developments affect word form realization, presenting a series of analyses of verb and noun articulations in conversational English. The analyses aim to determine the extent to which more permanent, grammatical constraints to the co-variate structure affect verb production, by contrasting verb and noun production. Do differences in the covariate structure predict the way younger and older speakers of English articulate nouns and verbs?

6.4 Analyses: Acoustic-Phonetic Deviation as a Function of Speaker Age and Uncertainty

Corpus Data. The Buckeye corpus contains phonetically transcribed speech from informal interviews with 40 speakers from Columbus, Ohio. The speakers are balanced by age and gender, the young cohort consists of speakers below the age of 30, the old speakers are aged 40 and upwards, with more age variance in the older speaker group. The 286982 words are annotated with a set of 41 ARPABET[16] labels expanded by a set of markers for the manner of articulation. The data set used in this analysis was extended by the duration of the silent pause preceding the words. For each segment, we added word and utterance position, aligned the observed and the dictionary form, and added a number of distinct lexical contexts (lexical bigrams) each word appears in in the Buckeye Corpus. We center and scale the highly skewed word and utterance position by part of speech to obtain a Gaussian distribution. The frequency counts for each form by part of speech were taken from an 80-million token subcorpus of contemporary American English (Davies, 2010) extracted from transcripts of unscripted conversation on TV and radio programs recorded between 1990 and 2000.

Segmental deviation. To estimate deviation in relation to word position we align the transcribed form to the dictionary form by mapping the absence or presence of the phonetic contrast presupposed by the dictionary model at its 'expected' position. Contrasts not found at their 'expected' position are marked with a special character. For example, the word *practitioner* presupposes the dictionary form *p r ae k t ih sh ah n er* and the observed form *p r ah k t ih sh n ah* is aligned as *p:p r:r ae:ah k:k t:t ih:ih sh:sh ah:_ n:n er:ah* so that the sequence position will not shift when contrasts in earlier positions are absent. The *deviation* at the centered and shifted indices *0.83 1.19 1.46 1.67 1.84 1.99 2.12 2.23 2.33 2.42* is then coded as *FALSE FALSE TRUE FALSE FALSE FALSE FALSE TRUE FALSE TRUE*. We do not take the label assigned by the annotator to necessarily correspond to the signal produced by the speaker, it merely suggests that the observed form is different enough to warrant alternative labeling.

Pause duration. We extract the duration of the silent pause preceding the word by calculating the time step between the words' onset and the previous words' offset. Note that in contrast to the analyses presented in chapter 5 the 'pauses' now include silences and audible breaths, puffs and coughs. Only 20% of the words are preceded by a measurable interruption (12.4% of nouns and 8.4% of verbs). Pause

[16] a phonetic transcription code developed by the Advanced Research Projects Agency (ARPA) to describe phonemes and allophones of American English, the two characters per segment system used here is the more common one

duration was log-transformed to approach a Gaussian distribution. The distribution has long tails, the left tail holds a bulk of pauses shorter than 180 ms, and the long right tail contains longer pauses that appear at utterance boundaries. Plotted separately, the 700 (of 9396) data points behind the right tail seem to approach a Gaussian distribution again. Thus it seems that pause distributions are aggregates of pauses from different ranges of duration.

The generalized additive mixed-effects model is a structure-exploration tool that can be useful in uncovering nonlinear effects of numerical covariates (Wood, 2017; Baayen and Linke, 2020). It is essentially an extension of the generalized linear model. In contrast to (generalized) linear models, which fit multivariate interactions with polynomial functions (weighted sums of the covariates), the GAM introduces smoothing splines, which allow fitting wiggly curves (combined smooth basis functions, such as, for example, cubic splines) to examine non-linear relationships between a response variable and multiple numerical and/or categorial predictors. Splines can take multiple arguments to model wiggly surfaces, distinguishing between factor levels (groups, categories, or experimental conditions). The differences between the surfaces can be visualized to inspect whether and where they differ significantly from each other. GAMMs introduce a weighted penalization method that aims to leverage between finding a good fit to the data and over-fitting to noise. In addition, the smoothing functions allow for constraining the dimensionality (number of basis functions or wiggliness of the effect) by setting the basis function parameter (k) to a fixed value below the default (10). One noteworthy limitation of the model is that it allows for the modeling of complex multidimensional interactions and random effects that provide satisfactory fits but interact with each other and the research question in uninterpretable ways. This makes it possible to fit models to effects (i.e., paint with GAMMs). Using GAMMs to gain an understanding of meaningful functional relationships between variables in structured data requires a detailed understanding of the model, its implementation, and a careful assessment of how both interact with the structure of the data set **and the research question**. In addition, when modeling linguistic variables, which tend to be correlated in ways relevant to research questions, GAMMs require paying close attention to auto-correlation and concurvity (the non-linear version of co-linearity) and how these phenomena affect the models' predictions (cf. Baayen and Linke, 2020). That is, the uncertainties involved in interpreting the model often entail research questions in themselves.

6.4.1 Preliminaries: Pause and utterance position as context

To examine how word form realization develops with speaker experience, we first define an independent variable to represent the 'context'. Results from chapter 5 indicate that variances in pause duration are independent of where and when they

occur (i.e., pauses seem to co-vary systematically with sequence length and are independent from individual differences in the distribution of words from different lexical categories and speaker experiences). Verbs and nouns are distinguished from each other by systematic patterns of co-variation with their lexical neighbors. As we have seen earlier, the differences at which uncertainty increases across word categories suggest that the relationships between words from different categories will change over time. The variability of (and the uncertainty over) verb argument frames will increase in relation to the verb argument frame (on average). The variability of nouns will increase in relation to the argument frame. The asymmetries in the distribution of information these developments entail suggest that the lexical context defined by transitional probabilities between words and their lexical neighbors is in constant flux. To sidestep this, we use pause duration as an independent estimate of contextual uncertainty. Pause duration approximates the information provided by the articulated prior. Because of the way pauses are distributed, very frequent, short pauses contribute little information on their own, and the information provided by the preceding part of the signal is preserved (i.e., the contextual uncertainty about the articulated signal[17] is low). Pauses longer than 250 ms, by contrast, are informative and modulate the information provided by the signal context. Signals following pauses longer than 3000 ms ought to be context independent because they can be assumed to occur outside the relevant temporal integration frame (White, 2017; Pöppel, 1997; Montemayor and Wittmann, 2014).

6.4.2 Are pauses functionally different from fillers?

To test this idea, we compare the relative probability of pauses and fillers preceding words from different lexical categories (verbs, nouns, and function words). Fillers and other auditory oddballs are functional in discourse: they draw attention to the deviant signal and anything that occurs immediately after it. This increase in attention, in turn, benefits recall. However, recent experimental results suggest that this function is restricted to later utterance positions, showing that only disfluencies in utterance-final positions elicit recall effects (Diachek and Brown-Schmidt, 2022). Moreover, speakers adjust their expectations about the upcoming part of the message only when disfluencies are perceived as 'natural', both non-native fillers and artificially manipulated pause durations fail to improve the perceived fluency and elicit recall effects (Bosker et al., 2014b; Cooke et al., 2014). In other words, the communicative function of disfluencies seems to be bounded by expectations (they ought to vary with experience and context). To count as signals, disfluencies need to be sufficiently expected in the context they occur in.

[17] Note that we are talking about sublexical variation here, where the realization of word-final and word-initial contrast(s) in connected speech generally tends to co-vary consistently with both the preceding and the following part of the signal. This effect (carryover and anticipatory co-articulation) becomes increasingly unlikely as the duration of the intermittent pause increases.

Although all disfluencies (fillers, repairs, and manipulated pause durations) appear to focus attention and benefit recall in experiments, there are several indicators that speakers perceive silent and filled pauses differently. Fillers seem to take a function comparable to that of morphemes or articles[18] (cf. Brennan and Williams, 1995; Kirjavainen et al., 2022). Suppose learning leads to a drift in the distribution of information across all articulated parts of the signal, and articulations vary in response to uncertainty. In that case, the fillers will, by definition, be affected by experience in ways pauses are not. The idea of the silence/sound dichotomy gains supports from a class of effects known under the term *phonemic restoration*.

Phonemic restoration is a perceptual illusion in which speech sounds that have been replaced by a cough or noise are perceived as distinctly pronounced and present in the utterance. The cough or noise themselves are localized elsewhere in the utterance by the subjects. The restorations can involve up to two or three phonemes (Warren, 1970), and the size of the restored sequence increases with the amount of information provided by the sentence context (Warren and Sherman, 1974; Samuel, 2001; Sivonen et al., 2006). The effect is generally discussed in terms of linguistic content (Groppe et al., 2010). It can, however, be evoked by speech-like rhythms even when actual sensory stimuli are absent (Cervantes Constantino and Simon, 2017). Older speakers seem to benefit more from this effect than younger speakers, especially when other sources of information are degraded (e.g., in vocoded speech and in interrupted sequences) (Saija et al., 2014; Jaekel et al., 2018). Notably, **the effect does not extend to silent segments**. Phonemes replaced by silences are reported missing and correctly localized by subjects.

The results summarized above suggest that phonemes and morphemes are replaceable in context and that the extent to which phonemes are replaceable will vary with the amount of information provided by the context. The findings summarized in earlier chapters suggest that older people benefit more from context (i.e., are less uncertain about the structure), and, consistent with this, older people are more likely to exploit the effects related to phonemic restoration (i.e., guess correctly).

Similarly, as speakers' uncertainty about the syntactic structure decreases, the probability of guessing correctly **in context** ought to increase in verbs and decrease in nouns. Any experience-related variability in the verb is increasingly likely to reduce the uncertainty about the surrounding context, while variability in nouns will become increasingly likely to reduce the uncertainty about the local competition (which nouns are not meant). Any remaining uncertainty (in English, at least) must, by definition, be distributed across the word (through hyper- or hypo-articulation) or its surrounding context (through filler insertion).

[18] many thanks to M.Ramscar for pointing this out, in personal communication

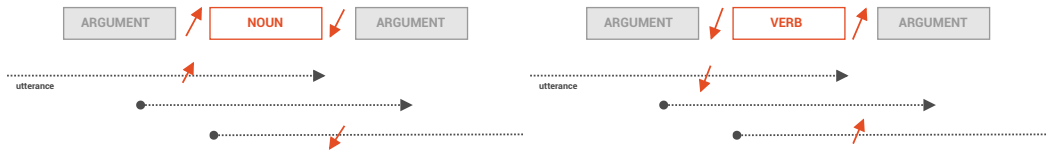

Fig. 6.4: Schematic illustration of the predicted uncertainty shift (i.e., misalignment in expectations) at the word-initial boundary of verbs and nouns in different utterance positions. The bottom part of the figure is meant to illustrate that, as the asymmetries between the lexical categories increase, the functional load on the verb to provide context (i.e., reduce uncertainty about its arguments) increases with utterance position, while the functional load on the noun to discriminate itself from competition decreases in later utterance position.

We have stressed at various points in this and the foregoing chapter that word learning continues across the lifespan. This suggests that because of the way words are distributed in the samples, speakers learn from[19], quantities estimated by metrics such as *surprisal* and *information content* (Hale, 2001; Demberg and Keller, 2008; Levy, 2008) will not be stable in time. **The asymmetries in the distribution of words from different categories in the cumulative samples leads to a following hypothesis**: the variance in the transitional probabilities between verbs and the lexical arguments they follow will decrease (and with it, the uncertainty at the transitions). Meanwhile, the variance in the transitional probabilities at the verb-final boundaries (between verbs and the arguments they predict), and at the transitions between nouns and the lexical frames that precede them, will increase [20]. The effect is illustrated in Fig. 6.4. This suggests that the uncertainty at the word-initial boundary should vary predictably with the lexical category, the relative utterance position, and the speaker's age.

To test this idea, we model the likelihood of filler and pause duration as a function of log utterance position and cohort (speaker age) for nouns, verbs, and function words separately. To further account for contextual uncertainty, we add a term for the following part of speech as a random effect.

```
Pause ~ s(logUtterancePosition, by = Cohort, k = 3) +
        s(Next POS, bs = "re"),
        data = verbs/nouns/function words
        family = "binomial"
```

[19] because word recurrence patterns are bursty.

[20] Note that we make strong simplifying assumptions about the 'model' utterance structure of English language, that will certainly not hold in many cases. Our focus is on the **distribution of uncertainty in the aggregate**, and explicitly not on single-point estimates or exemplars. Individual instances will, of course, vary in ways that will not conform with the aggregate model.

In pause duration, the utterance position and speaker age explain 70% of the variability in pause duration. The effect is u-shaped, and most of the variance in pause durations is found at the boundaries of longer utterances (Figure 6.7). Older speakers produce shorter pauses in utterance final positions, independent of the lexical category. There are no differences between verbs, nouns, and function words and the upcoming part of speech. This effect is consistent with other findings that also suggest that individual variation in articulation rate is an epiphenomenon of phrase length and disappears when phrase length is controlled for in the analysis (Crystal and House, 1990; Quené, 2005).

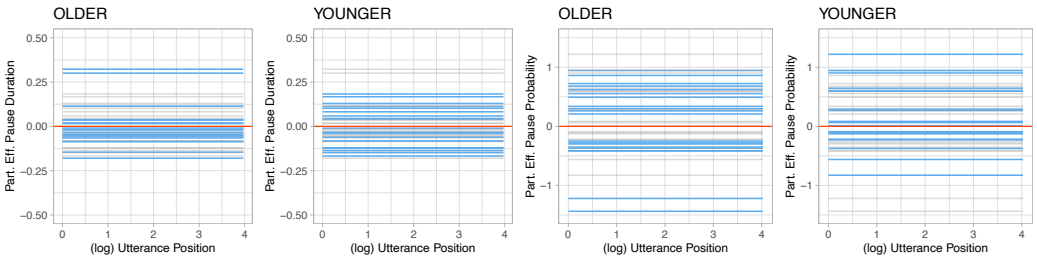

Fig. 6.5: Factor smooths for individual speakers from different age cohorts. The target cohort is highlighted in blue. Left: Factor smooths for pause duration as a smooth function of utterance position. Right: Factor smooths fitted to the log odds of silent pause for utterance position in older speakers and younger speakers. Variation in pause duration and pause likelihood seems to be distributed across utterance positions of individual speakers (variance in utterance length increases in older speakers, see Tab. 8.3, in the appendix), while individual variation in pause production decreases). Younger speakers' smooths are distributed symmetrically around the intercept. In older speakers, there are more outliers, while the bulk of speakers is distributed densely around the intercept. The between-speaker variability seems to be explained by the variation in utterance length. For comparison with articulation, see Fig. 8.3 in the Appendix

By contrast, the log odds of fillers vary with the cohort and the uncertainty of the upcoming word category. This finding is consistent with the suggestion that experience affects the distribution of information across all parts of articulated signals.

```
Filler ~ s(logUtterancePosition, by = Cohort, k = 3) +
         s(Next POS, bs = "re"),
         data = verbs/nouns/function words
         family = "binomial"
```

In the next part of the analysis, we contrast the relationship between pause duration and predictability parameters for verbs and nouns; Does the relationship between lexical uncertainty and pause duration change with speaker experience? We then examine the deviation from the dictionary form against pause duration and word/ut-

Probability of Pause, nouns				
A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	-4.3000	0.0465	-92.4262	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
s(Utt. Position):older	1.9996	2.0000	4881.2805	< 0.0001
s(Utt. Position):younger	1.9995	2.0000	4464.6914	< 0.0001
s(Next PoS)	2.6835	41.0000	3.5436	0.1862
Probability of Pause, verbs				
A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	-3.6857	0.0394	-93.5811	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
s(Utt. Position):older	1.9996	2.0000	5974.5313	< 0.0001
s(Utt. Position):younger	1.9996	2.0000	5656.6848	< 0.0001
s(Next PoS)	2.4171	43.0000	3.5103	0.1296
Probability of Pause, function words				
A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	-3.0087	0.0402	-74.8562	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
s(Utt. Position):older	1.9999	2.0000	17361.3415	< 0.0001
s(Utt. Position):younger	1.9999	2.0000	15960.4015	< 0.0001
s(Next POS)	19.4850	43.0000	122.8819	< 0.0001

Tab. 6.1: Pause Probability Model - log odds of pause preceding verbs/nouns and function words as a function of smooth over (log) utterance position, and speaker cohort, part of speech of the word following the verb/noun/function word is added as a 'random effect' (note that there is no significant random effect for parts of speech following verbs and nouns.

terance position: does the relationship change with experience? Do nouns and verbs differ in the extent to which this is the case?

6.5 Tests of the main hypothesis

Results from the previous section support the idea that the relationship between articulations from distinct categories and pause duration and utterance length (and position) develop in a predictable way across the lifespan. We proceed with an exploration of the relationship between pause and position (of words in utterances, and phonemes in words), and the way position and pause affect form articulation. Our hypothesis is that the likelihood of deviation from the dictionary form will increase in signal dimensions that show convergent patterns of development in older speakers (forms that are predictable) when contextual support is weaker (at the utterance initial boundary, following longer pauses in later positions in the word). By contrast, articulated signals that become increasingly diversified across the lifespan (e.g., word initial contrast in verbs and nouns), ought to resemble the dictionary form, which,

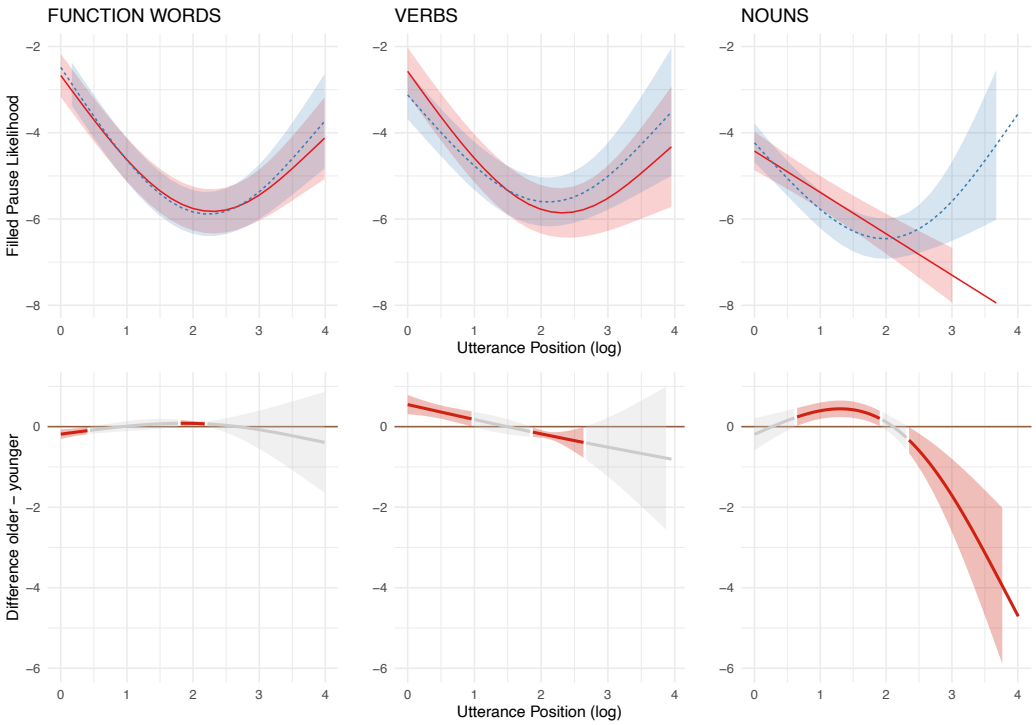

Fig. 6.6: Top: Likelihood (log odds) of filled pause as a smooth function of utterance position in older speakers (red) and younger speakers (blue). Bottom: The difference curve of filled pause odds for older and for younger speakers. Parts of the curve that have 95% confidence intervals that do not include the horizontal line are highlighted in red, they differ significantly from zero. We can see that older speakers are less likely to articulate nouns preceded by fillers in later utterance positions. The differences between the cohorts seem to increase with the lexical productivity of the class and utterance position. These results support the idea that the inconsistency in speakers' estimates of conditional probabilities between functors and content words will increase across the lifespan. This seems to suggest that conditional probabilities between words in a corpus cannot provide reliable estimates of contextual uncertainty across individual speakers.

as we have suggested approximates a signal that provides maximal discriminability when no context is provided (in isolation).

We also predict that this will affect nouns in the way it will not affect verbs in older speakers. This prediction is based on the assumption that the functional load on the discriminative contrasts at the word boundary increases with contextual uncertainty and the uncertainty related to the class of words the context predicts (which, as we have discussed earlier, can shift across the lifespan). We discussed earlier that older speakers uncertainty about nouns ought to increase across adulthood. Therefore, we expect the noun initial boundary to increase across the lifespan as the uncertainty increases.

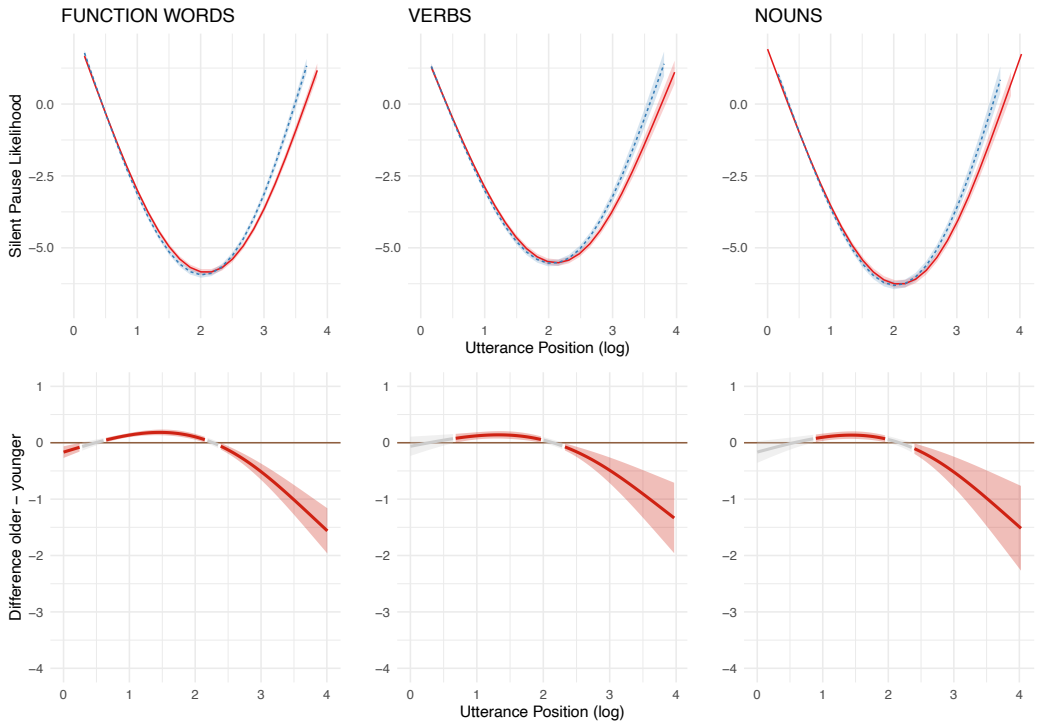

Fig. 6.7: Top: Likelihood (log odds) of silent pause as a smooth function of utterance position in older speakers (red) and younger speakers (blue). Bottom: The difference curve of silent pause odds for older and for younger speakers. Parts of the curve that have 95% confidence intervals that do not include the horizontal line are highlighted in red, they differ significantly from zero. Older speakers are less likely to articulate words preceded by pauses in later utterance positions, independent of the words' lexical class. These results support the idea that pauses can serve as reference points to measure acoustic variation across individual speakers.

6.5.1 The Effects of Utterance Position, Collocate Diversity and Frequency on Pause Duration

We model the duration of the preceding pause as a function of collocate diversity, log frequency, and utterance position, adding lexical category as a co-variate term to all three continuous predictors. Our data set comprises all verbs and nouns preceded by a pause, 5618 of the 44257 noun tokens, and 3778 of the 44340 verb tokens in the Buckeye corpus. To account for non-linearity and individual differences, we model the interactions as smooths with a generalized additive mixed effects model (Wood, 2017), adding individual speakers as a random effect. The model is specified as follows

```
PauseDuration ~ s(logUtterancePosition, by = POS, k = 3) +
                s(CollocateDiversity, by = POS, k = 3) +
                s(logFrequency, by = POS, k = 3) +
                s(Speaker, bs = "re"),
```

```
method = "ML"
```

In the exploratory analysis, we find no significant differences in pause duration between younger and older speakers. If the relationship between the pause and the utterance position changes across the lifespan, the effect seems to be related to changes in utterance length and the rate at which words from different frequency registers occur (e.g., the way older speakers use nouns in utterances). In line with analyses in chapter 5, pauses are not affected by experience, instead, the utterance structure (the way words are used in sequences) appears to change consistently in relation to pause.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	-0.8275	0.0381	-21.6935	< 0.0001
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(UtterancePosition):noun	1.9638	1.9986	16.6048	< 0.0001
s(UtterancePosition):verb	1.9256	1.9943	9.0411	0.0001
s(CollocateDiversity):noun	1.9865	1.9997	193.6850	< 0.0001
s(CollocateDiversity):verb	1.1750	1.3191	0.5044	0.6447
s(logFrequency):noun	1.9718	1.9991	114.9864	< 0.0001
s(logFrequency):verb	1.0006	1.0012	1.4144	0.2345
s(Speaker)	34.6189	39.0000	8.1325	< 0.0001

Tab. 6.2: Summary of **model 0** – pause duration as a function of (centered) utterance position, collocate diversity, and word frequency with separate smooths for verbs and nouns, and a by-speaker random effects.

We observe effects of utterance position, collocate diversity, and log frequency in nouns (6.2), and a fair amount of variance in individual speakers ($edf = 34.619$, $p < 0.0001$). In contrast, we find no effects in verbs. Utterance position yields a low p-value in nouns and verbs (the slopes are significantly different from zero). The effect appears largely driven by outliers found at the boundaries of very long utterances (pauses preceding words at the boundaries of very long utterances are disproportionately shorter in the model). The effect of position on the duration of pauses at the initial boundary of shorter (and most frequent) utterances does not seem to be significantly different from 0 (see 6.8, panel B). The gray density area at the bottom of the plot represents the distribution of utterance positions in all verbs/nouns, and the area under the red density line represents the distribution of verbs/nouns preceded by pauses. The difference between the densities reveals that most verbs and nouns preceded by pauses occur early in short utterances. This, in turn, supports the idea that silent segments are independent signals that modify the information rates in the sentence context (i.e., they occur in bursts where needed). The management of expectations appears to be achieved through the joint contribution of articulated signals, which modify the rate by incrementally decreasing the variable error (segmentation rate variability), and silent signals, which reset the expectations about the variable error (variability of timings in speech sequences).

Collocate diversity and log frequency are highly correlated in both nouns ($r(5551) = 0.777$, $p < 0.0001$) and verbs ($r(3773) = 0.941$, $p < 0.0001$), and both values yield high scores when we test for concurvity in the model (verbs 0.88 and 0.89, and nouns 0.65 and 0.63). Concurvity scores (a value ranging between 0 and 1) measure over-specification in the model. It is a generalization of co-linearity, an effect that arises when highly correlated predictors compete for effects in sparse data. In these cases, it is hard to tell which of the variables is driving the model prediction and whether the model prediction can tell us something about the relationship between the variable and the response (see Baayen and Linke, 2020). In this model, however, this does not seem to be a problem. Both measures seem equally uninformative in predicting the duration of pauses preceding verbs. The model yields stable effects in nouns that reveal differences between the recurrence phenomena collocate diversity and corpus frequency capture[21]. Model 0 indicates that pause duration interacts with noun predictability, but not with verb predictability. The model predicts opposite effects of diversity and frequency in nouns: low collocate diversity predicts shorter pauses, and low frequency predicts longer pauses. Pause duration increases with collocate diversity and decreases with frequency. The misalignment between the two predictors may reflect the burstier recurrence patterns that nouns generally have and the effect on the uncertainty preceding nouns. How?

The relationship between burstiness, the local increase in the relative frequency of low-frequency nouns in certain contexts, and a decrease in collocate diversity can be understood as follows: locally, both utterance context and discourse structure can facilitate the predictability of low-frequency nouns. Introducing a new word in a conversation usually takes a specific lexical context, an utterance. Once mentioned, the low-frequency word can be used without further introduction. It is neither new nor unexpected in the context of the unfolding interaction. The word *distribution*, for example, is a low-frequency word. The Buckeye corpus does not contain the word *distribution*. In chapter 3, by contrast, the word distribution occurs 151 times; 36 times preceded by the definite article *the*. This frequent co-occurrence of nouns and functors reduces the collocate diversity of nouns in contexts in which they are expected. Thus the contrast we observe in the frequency and collocate diversity effect likely reflects the degree to which the uncertainties and information rates are managed by pauses and functors in utterances (cf. Frantzi and Ramscar, 2022; Brevi and Ramscar, 2022). The idea may at first seem a bit puzzling and counter-intuitive:

[21] Nouns that are infrequent in the larger corpus and co-occur with multiple lexical neighbors in the smaller corpus are over-represented in the smaller corpus (they are context-specific to the corpus).

earlier, we stressed that pauses and articulations are functionally distinct[22], how are pauses related to functors?

The is one of the most frequent words in English. It is more frequent in text than in speech, presumably because speakers rely on other signals to achieve similar communicative results. *The*, in text, contributes little in terms of uncertainty reduction about the messages' meaning. Instead, it structures the utterance for efficient transmission and manages the unfolding information rates; it is a functional equivalent of fillers, which, as we have suggested earlier, appear to provide information about the relative uncertainty of the word in context and help guide speakers' attention. Low-frequency *distribution* co-occurs with a distribution of collocates. The word 'distribution' modified by *the* (36 times), *geometric* (8 times), *power-law* (7 times), *empirical* (6 times), etc., represents different uncertainty registers mediated by the grammatical context (*determiner+noun* vs. determiner+adjective+noun) (cf. Dye et al., 2018; Levy, 2008). The grammatical contexts, as we have shown earlier, follow their own distributions in speech sequences (see Chapter 3, Figure 3.6) and reflect uncertainties managed at the more abstract level of utterance length and the relative utterance position.

In other words, many different factors, such as situational contexts, world knowledge, grammar, and discourse structures, determine the uncertainty of lexical contexts in which words occur. Speakers implicitly respond to this uncertainty by adapting the structure of utterances they produce in ways that maintain the systematicity of distributions. This suggests that despite the complex interactions that shape the ebb and flow of uncertainty in communication, the individual contributions at the transition between any two words can be simplified in terms of the relative sequence position and the relative uncertainty of rates represented by a pause duration (see also chapter 5).

6.5.2 Uncertainty and Variation in Speech over Time

In the first model, we operationalize misalignment between the transcribed word initial segment and the dictionary form as an indicator of functional load on the discriminative contrast.

[22] Note that function, as we define it here, unfolds over a continuum. The assumption is that articulated signals can change in ways pauses cannot. The degree to which a functor is 'segmented' will vary. 'Function words' and particles, such as articles or morphemes, can be segmented if the context requires or allows fine-grain contrast. At points where *information contents* fluctuate with experience (i.e., at the utterance boundaries), functors ought to become **increasingly discriminated or increasingly indiscriminate** as speakers' experience grows. Articles at the word-initial boundary ought to become increasingly 'smudged,' while articles at the word-final boundary ought to become distinctly 'articulated' (in English, at least).

Word initial boundaries are the least likely site of deviation from dictionary form in both verbs and nouns. In connected speech, the likelihood of encountering speech contrast presupposed by the dictionary form decreases with the word position. The word-initial *deviations* comprise only 4% of the data points in nouns and 9% of verbs (compared to 18% of adverbs and 51% of function words). Further, only 18% of nouns and 6% of verbs that depart in word-initial contrast are actually preceded by a pause, 77% of them are found the phrase initial boundary.

Model 1, nouns: We analyzed differences in the realization of 5032 noun initial phones as a function of utterance position and the duration of preceding pause by fitting a tensor product smooths for each cohort. Figure 6.9 illustrates a three-way interaction between the two covariates and two factor levels for older speakers (row 1) and younger speakers (row 2). We obtain two hypersurfaces (Figure 6.9, right column), both relatively wiggly. We find no interpretable effects in younger speakers' data, the confidence intervals are wide and overlap. Initial segment deviation does not seem to interact with pause duration or utterance position in younger speakers.

The surface fitted to older speakers' data reveals an interaction between pause duration and utterance position; the likelihood of deviation increases with utterance position as pauses get longer and decreases with utterance position when as pauses get shorter (see 6.9, right column, row 1). Figure 6.9, left column, row 3, shows the difference between the hyper-surfaces fitted to older and younger speakers data. The color coding in the contour plot represents the magnitude of difference when older younger speakers' surface is subtracted from the older speakers' surface. Older speakers seem to become increasingly sensitive to pause duration, the likelihood of word-initial deviation decreases with pause duration in pauses longer than 500 ms and decreases in pauses shorter than 500 ms. The results indicate that older speakers' articulations reflect sensitivity to utterance structure and timing in nouns, while younger speakers' articulations do not.

Model 1, verbs: In older speakers, the log odds of deviation increase with pause duration and utterance position. The effect is moderate and limited to the bulk of data in the center of the plot. Given that we are looking at the interactions between scaled variables, the effect does not allow for many certainties. In younger speakers, we again find no interpretable effects.

Our results suggest that older speakers' articulations become increasingly adapted to the context in nouns (where the uncertainty increases across the lifespan) but not in verbs. In younger speakers, we find no effects. To explore whether these results replicate across word positions, we fit a second model to assess the log odds of deviation as a three-way interaction of pause duration, segment position in the word, and speaker age. The model is specified as follows:

NOUN MODEL 1				
A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	-2.542	0.057	-44.931	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
te(Utt. Pos.:Pause Dur.):older	3.000	3.000	50.462	< 0.0001
te(Utt. Pos.:Pause Dur.):younger	8.497	10.526	88.526	< 0.0001

VERB MODEL 1				
A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	-2.442	0.064	-38.193	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
te(Utt. Pos.:Pause Dur.):older	5.209	6.345	17.062	0.011
te(Utt. Pos.:Pause Dur.):younger	7.819	9.870	25.140	0.005

NOUN MODEL 2				
A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	-1.471	0.041	-35.535	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
te(Segment Pos., Pause Dur.):older	10.316	12.390	394.161	< 0.0001
te(Segment Pos., Pause Dur.):younger	11.328	13.433	465.762	< 0.0001
s(Speaker)	31.614	39.000	190.649	< 0.0001

VERB MODEL 2				
A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	-1.475	0.055	-26.703	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
te(Segment Pos., Pause Dur.):older	5.757	6.434	310.272	< 0.0001
te(Segment Pos., Pause Dur.):younger	6.256	7.511	254.516	< 0.0001
s(Speaker)	31.247	39.000	180.164	< 0.0001

Tab. 6.3: Summaries of **model 1**, that fits the log odds of deviation as a three-way interaction of the duration of the preceding pause, utterance position, and speaker age, and model 2, that fits the log odds of deviation as a three-way interaction of the duration of the preceding pause, segment position in the word, and speaker age.

```
Deviation ~ te(PauseDuration, PhonePosition, by = Cohort) +
              s(Speaker, bs = "re"),
              family = "binomial"
```

The command *by=Cohort* fits one 3-dimensional surface for each age group. In Figure 6.10, the hyper-surface is visualized in a contour plot, where contour lines connect points with the same partial effect. The color coding represents magnitudes of the partial effect of preceding pause duration (y-axis), phone position in the word (x-axis), and speaker age (younger and older speakers, and the difference between older and younger speakers' surfaces).

Model 2, nouns: In nouns, the distance between the contour lines indicates that the effects are stronger in older than younger speakers. In younger speakers, the likelihood of deviation increases with segment position and pause duration. In older speakers, we find a more complex interaction. In pauses shorter than 2 seconds, the likelihood of deviation increases with segment position, independent of pause duration. Older speakers seem to be more likely to maintain word-initial and word-final contrast following longer pauses.

Model 2, verbs: In verbs, by contrast, the effects do not appear to get stronger. Instead, the effect seems to reflect a shift in the way speakers respond to uncertainty. Older speakers seem to be more likely to maintain word-initial and word-final contrast in verbs following shorter pauses, and younger speakers seem to be more likely to maintain verb-initial contrast following longer pauses. The pattern of variation itself does not change, only the magnitude and the resolution do. The likelihood of variation at both word boundaries decreases with experience and uncertainty.

Taken together, our results show that experience changes the way speakers articulate words and that this change affects nouns and verbs differently. Older speakers are more likely to deviate at the noun boundaries when the uncertainty is low and less likely to deviate when the uncertainty is high. In verbs, experience leads to an opposite effect, the likelihood of deviation at the verb boundary decreases when the uncertainty is low and increases when the uncertainty is high.

6.6 Discussion: Word form evolution across the lifespan

These findings seem to support our hypothesis that competition for acoustic contrast will increase more with speaker experience in nouns than it will in verbs. Model 0 shows that there is an interaction between pause and noun predictability and that there is no interaction between pause and verb predictability. In addition, model 1 reveals that the likelihood of deviation in older speakers interacts with pause duration and utterance position in nouns but not in verbs (or younger speakers).

In model 2, speaker experience seems to reverse the effect of pause duration (uncertainty) in word-initial positions of verbs; older speakers are less likely to deviate from the dictionary form following a short pause and more likely to deviate following a long pause. The uncertainty at the word-initial boundary (at the transition between the preceding part of the utterance/signal and the verb) is low. Words that are preceded by short pauses (shorter than 250 ms) are embedded in a well-rehearsed utterance structure, and articulations unfold habitually/fluently in 'bouts.' As we have

discussed in the previous chapter, pauses from this range of durations seem to contribute to the adjustment of temporal expectations and do not seem to lead to adaptation of speakers' expectations about changes in the acoustic dimensions (see section 5.2). Words preceded by a longer silent interval signal a disconnect between the articulated prior and the word. The pause disrupts the articulation dynamics, increases the uncertainty and focuses attention to the realization of the word initial segment. The likelihood of deviation decreases with uncertainty and as utterances get longer (when contextual support is high and as articulation rates increase). In other words, noun articulations co-vary with the utterance context and pause duration. This suggests that phoneme deviation is too coarse grained to fully examine the complex interaction, and that these findings can inform and motivate further analyses.

In contrast to the general tendency to deviate more in later word positions, older speakers seem to deviate less in word-final positions of longer verbs. These differences in the extent to which cohorts attenuate word-final boundaries in longer words may reveal an increasing sensitivity to functional aspects of verb morphology: more frequent and contextually dispersed verbs in English are far more likely to take shorter, often idiosyncratic forms than less frequent and less contextually dispersed verbs. Irregular verbs take explicitly lexicalized forms, rather than conforming to functional classes that share inflectional features. From the systemic perspective, this makes verbs that take irregular forms exempt from the functional pressures of the class, which in turn allows them to be more contextually promiscuous. Regular verb forms, by contrast, provide consistent (and informative) patterns of verb inflection that systematically interact with contexts (and, by consequence, different semantic interpretations, i.e., word meanings). Inflectional categories contribute little to reducing the uncertainty about the verbs (i.e., providing contrastive features to discriminate verbs from other verbs), instead, they decrease the uncertainty about the message context (how or when events occur, i.e., tense and aspect) (Bybee, 1985; Ramscar, 2002). Verbs marked for tense and aspect are not free to collocate without restraint with other arguments. This, in turn, means that verbs marked by regular patterns of inflection provide important information to reduce uncertainty about the upcoming word. The strengthening of word-final articulations could thus reflect an increase in sensitivity to this important functional role of verb endings in inflected verb forms. As discussed in chapter 3, nouns are not bounded by the grammatical constraints that utterances impose on regular verbs. Whether noun-final contrast is functional and necessary will vary much more with the context in which the noun appears (see e.g. Mahowald et al., 2013). Nouns preceded by shorter pauses receive more support from context. The fact that their final parts are realized consistently by older speakers following longer pauses, may thus reveal that speakers get more sensitive to the way sublexical components of nouns interact with context.

These observations and the questions and considerations their analyses entail, reveal some of the complexities and uncertainties involved in mapping models to processes. They also seem to underline the idea that convergence and consolidation of certain aspects of speech contrasts are as functional as the divergence and increase in contextual dispersion in others, and that these distinct responses jointly allow systems of communicative contrasts to adapt to functional pressures. They also seem to suggest that the functional dynamics behind speech variation cannot be interpreted meaningfully without accounting for the distinct functional roles sublexical contrasts at various levels of abstraction fulfill in communication.

6.7 Summary and conclusions

In this chapter we examined the conditions under which word forms realized in continuous speech signals deviate from word form models (dictionary forms). Our results provide support for functional variation in context: speakers adapt the articulated signals by increasing contrast in informative and decreasing contrast in the less informative parts of signals. In contrast to previous findings, we show that informativeness of distinct aspects of speech changes with speaker experience and that articulated signals develop predictably across the lifespan.

We presented evidence that differences between the cumulative verb and noun distributions in speech samples affect the way older and younger speakers articulate word forms in spontaneous speech. Verbs and nouns are lexical categories similarly represented by token count yet distinct in the number of individual types they host and the degree to which their collocate structure is fixed by grammar. To examine how differences in recurrence patterns and collocate diversity affect the resolution of speech signals across the speakers' lifespan, we analyzed experience-related changes in the articulation of verbs and nouns, comparing annotated speech sequences of older and younger speakers of English.

The 4-part-analysis provides evidence that all examined relationships between different aspects of articulated signals change with speaker experience. In the initial analyses, we tested the hypothesis by modeling the difference in the probability of filled and silent pauses in utterances produced by older and younger speakers. The results suggest that the likelihood of disfluencies and utterance position co-varies consistently with the utterance length. Disfluencies are more likely in initial positions and less likely in later positions in utterances produced by older speakers.

In contrast to silent pauses, where the effects are identical across word classes, the uncertainty associated with the word class seems to modify the effect in filled

pauses. The likelihood of fillers in final utterance positions decreases with speaker age, and the effect is more pronounced in nouns than in verbs. That is, experience-related uncertainty (i.e., the knowledge that there is more uncertainty associated with nouns than with verbs) seems to have an impact on articulations (including filled pauses) but not on silent pause production. In a model, we show that, in contrast to the articulated parts of signals, pause duration and pause likelihood appear to vary across speakers, reflecting differences in the length of utterances produced by individual speakers. Pause productions seem to vary with individual speakers' models, but do not seem to vary in the aggregate.

In the second part of the analysis, we show that pause duration co-varies with word frequency and contextual dispersion (n-gram diversity) in nouns but not in verbs. We find no differences between the cohorts. This supports the suggestion that the relationship between the pause duration and the relative predictability of the word will become more informative across adulthood in nouns (because the relative frequency of nouns will change). The third part of the analysis provides further support for this, showing that word-initial deviation in nouns seems to be a function of utterance position and pause duration and that older speakers model this relationship more consistently. Finally, we show that patterns of articulation in verbs and nouns develop differently across the lifespan.

These observations raise questions about the way learning and population structure contribute to language. They suggest that speaker experience develops across the lifespan and with it the speech signals speakers' produce. Shared codes and the temporal structure of signals are maintained and transmitted by speaker collectives. Does this imply that an increase in lifespan expectancy can lead to systematic changes in the way languages develop (which parts of the signal are maintained)? Do non-linear developments in the aggregate structures we observe in linguistic distributions reflect adaptation to rapidly diversifying signals? Does this perpetually fragment the signal structure?

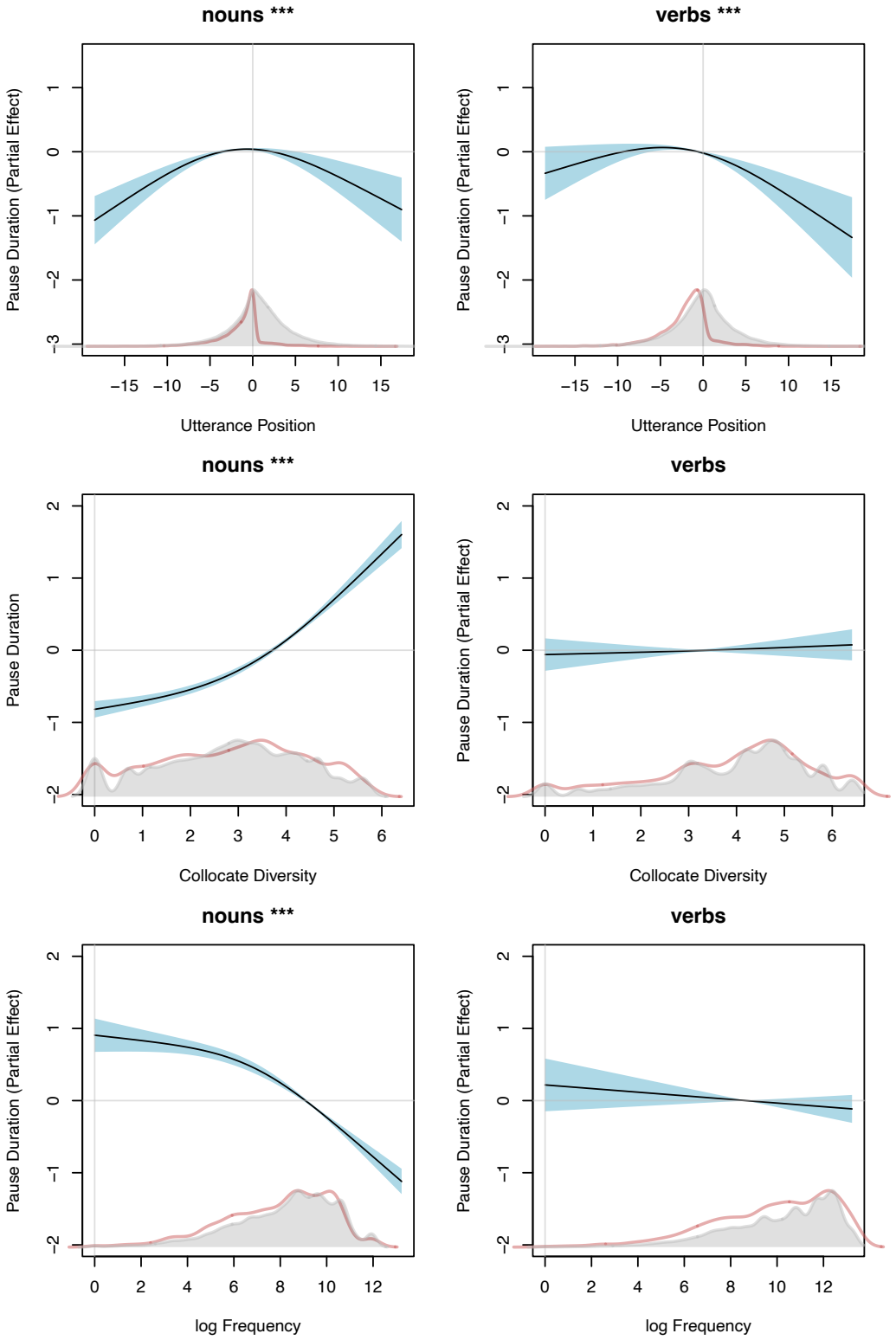

Fig. 6.8: Pause duration as a smooth function of utterance position (row 1), collocate diversity (row 2), and log frequency (row 3). nouns on the left side and verbs on the right side. The grey areas at the bottom line show the density of the distribution for all words in the category and the red line for words included in the analysis - all words preceded by a pause. The plots show that the duration of the pauses preceding nouns (left) decreases with frequency and the proximity of an utterance boundary and increases with the number of lexical contexts the word occurs within, while neither of the former appears to affect the pause duration preceding verbs (right column).

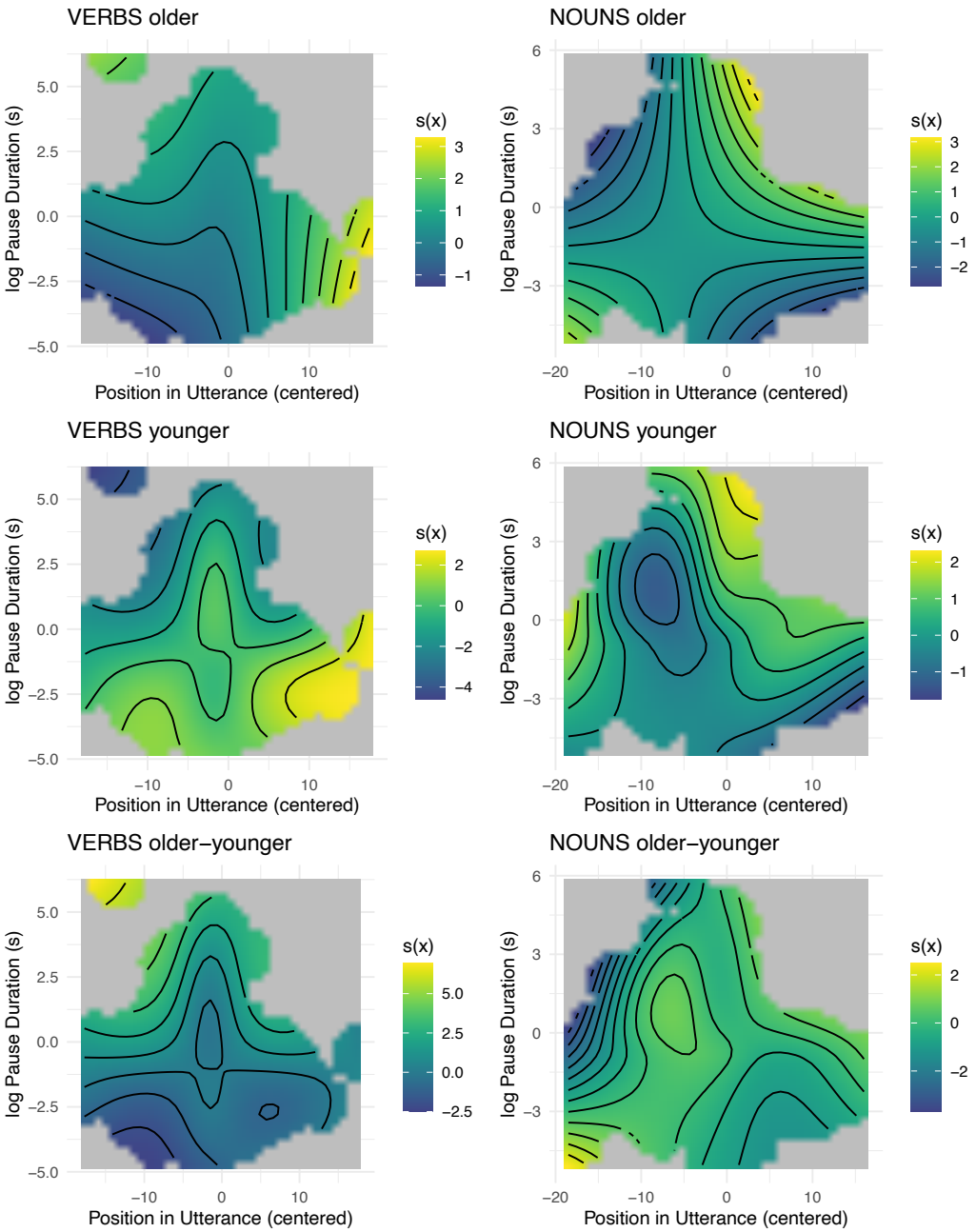

Fig. 6.9: Likelihood (log odds) of initial phone deviation (z-axis) as a three-way interaction of the preceding pause duration (y-axes), utterance position (centered, x-axes), and speaker age (top row - young speakers, center row - old speakers, bottom row - difference). Variance in word-initial contrast in nouns (right column) interacts with utterance position and pause duration in older speakers but not in younger speakers. There is no consistent effect of pause and utterance position on the deviation in verb-initial segments.

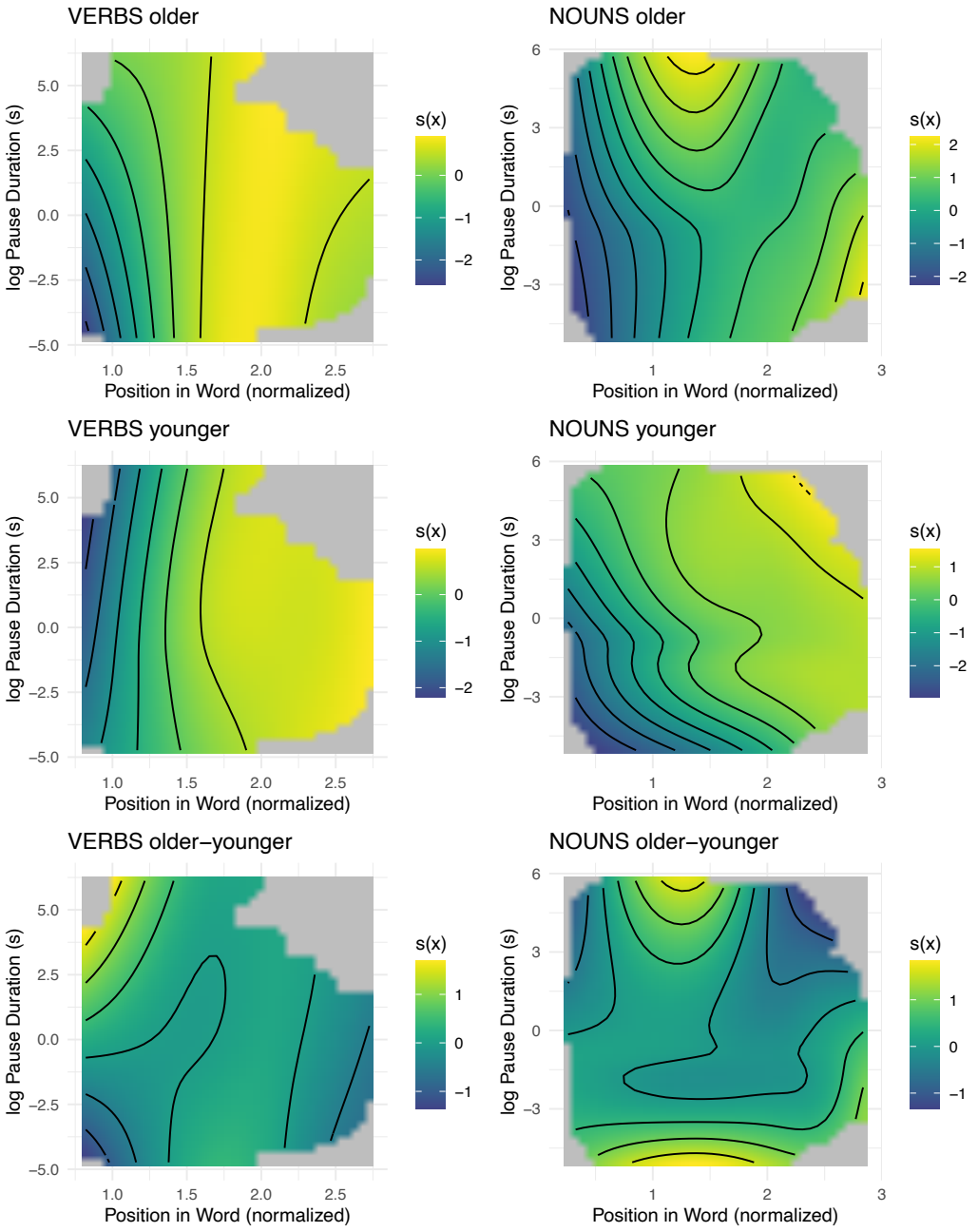

Fig. 6.10: Log odds of phoneme deviation (z-axis) as a three-way interaction of duration of preceding pause (y-axes), position in the word (x-axis), and speaker age (young speakers - top row, old speakers - center, difference old to young - bottom row). The left column shows verbs, and the right column nouns. The plots indicate that while in both verbs and nouns, the likelihood of deviation increases with word position, the interaction develops differently in relation to pause duration and word position across the lifespan. The model predicts opposite trends for final and non-final contrast in verbs and nouns preceded by longer vs. shorter pauses. The difference plots (bottom row) suggest that experience increases the difference in the articulation patterns of verbs and nouns in context.

7

Higher quality mules: How much analysis can we infer from data?

> *So we close with a Damascus joke. One vendor was selling the exact same variety of cucumbers at two different prices. "Why is this one twice the price?", the merchant was asked. "They came on higher quality mules" was the answer. – We only judge a technology by how it solves problems, not in what technological attributes it has.*
>
> — **Nassim Taleb**
> (Bitcoin, Currencies, and Bubbles)

The findings presented in previous chapters indicate that information contents of speech sequences are unstable quantities that can fluctuate with the communicative task, the uncertainty of the context, and individual speakers' experience. These findings seem to undermine many established beliefs about human behavior in general and human communicative behavior in particular. Many aspects of linguistic research rely on the existence of explicit, unambiguous forms in explaining the processing and transmission of information. We have, at various points in this thesis, stressed that linguistics 'forms' are cultural 'artifacts' supported by top-down constraints to the transmission process (i.e., centralized education and mass literacy) and the need to maintain mutual predictability. Our results suggest that many aspects of communicative codes are maintained through collective 'computations' and that languages, in their nature and structure, represent complex social systems. This characterization of human communication seems to bear on models of behavior at all levels of analysis. In what follows, we show how differences in the assumptions about representations, tasks, mechanisms, and their implementation can lead to very different conclusions based on unchanging evidence. The analysis illustrates that principles that shape and constrain human communication seem to apply at higher levels of analysis as well.

In this chapter, we illustrate some of the problems of overdetermined models[1] in an analysis of materials from a series of widely acknowledged comparative studies on human-animal cognition reported by Grainger et al. (2012) and Ziegler et al. (2013). Comparative studies examine similarities and differences between human and non-human behavior (e.g., the structure of vocal communication in humans and other animals). The idea behind these studies is that behavioral differences can help reveal the mechanisms involved and explain the conditions under which complex abilities unique to our kind arise. Some of these studies use linguistic materials or other 'cultural artifacts' to asses animals' ability to achieve human-like performance. Our analyses here suggest that many of these studies rely on questionable assumptions about the nature of both human and animal performance.

Nevertheless, the animals' performance in these experiments raises many interesting and challenging fundamental questions: How much can be learned from input under different conditions? Which features of complex behaviors are owed to specific neuro-biological features? Are complex, distributed neurophysiologies required to implement mechanisms that allow complex mental computations? Are complex computations needed to solve problems in experimental tasks? What exactly are people (and animals, and yet other people) doing when they are problem-solving? These questions tap into a number of 'difficult' topics left unanswered by theoretical and empirical investigations of intelligence, decision-making, structure, and representation. To address some of them, we present a simple, quantitative exercise, that reveals difficulties involved in distinguish behaviors guided by complex top-down processes from seemingly less sophisticated, automatized behaviors.

7.1 The curious case of reading baboons

Comparative studies have been used to shed light on various aspects of language, including reading. These studies typically examine what animals can learn, and pay less attention to how this interacts with the structure of what they learn from. We present a distributional analysis of the materials employed in baboon reading experiments, and, in relation to a learning simulation, show how aspects of the human-like behavior exhibited by non-human primates can be explained by the latent structure of the linguistic materials used in their training. We show how animal performance in a reading task, along with the levels of representation required to achieve this performance, depends on the specific task and the specific structure

[1] e.g., models that have more parameters than questions to answer. In mathematics, an **overdetermined system** of equations is a system in which there are more equations than unknowns. Overdetermined systems can provide solutions for special cases or cases where solutions are linear combinations of others (new models are combinations of old models). Overdetermined systems tend to be inconsistent and contain equations that do not contribute to the solution set of the system.

of the materials used in training and testing. These results indicate that a better understanding of the contribution of animal models may require a more nuanced approach than asking simple learnability questions.

7.2 What do animal models tell us about reading in humans?

Human communicative skills are unparalleled elsewhere in the animal kingdom. This raises challenges when it comes to explaining the origins of these otherwise unprecedented capacities. Many of these challenges relate to two closely connected questions: which of the representations and processes that underlie our linguistic capacities are learned by individuals, and which of these have been learned collectively, through evolution? An obvious way of investigating the demarcation between individual learning and biological evolution in language acquisition is to use animal models. Humans share their basic learning hardware with numerous other species, and so, in theory at least, testing to see whether a posited linguistic representation and / or process can be acquired by another species (even which species) can shed light on whether a given aspect of linguistic cognition is human specific – and presumably evolved – or not (while comparisons between other species offer the possibility of establishing the kind of neural mechanism required to learn it).

This approach may not have always covered itself in glory when it comes to language tout court (Kulick, 2017), yet it has proven useful when considering matters of nature and nurture in relation to specific aspects of the communicative process, such as in comparative studies of the ability of humans and animals such as chinchillas to perceive speech sounds (Kuhl, 1981). Another aspect of the human communicative repertoire where this approach has proven fruitful is in the study of reading. The human capacity for reading developed only in the last 5,000 or so years, making it a relatively recent addition to human communicative skill (by contrast, by some estimates spoken language emerged some 400,000 years ago; (De Boer, 2017)). Given that this indicates that the representations and processes associated with reading are unlikely to have evolved biologically, it raises a number of empirical questions about their learning that are amenable to comparative study.

Comparative studies of reading To this end, studies have shown that baboons are capable of discriminating between actual words (as defined in their training schedule) from non-words. Moreover, like humans, baboons are also sensitive to the similarities between English 4-letter words and random character sequences of same length, classifying orthographically similar strings as words, and dissimilar strings as non-words. These findings have been taken to indicate that the baboons,

who share more neurophysiological features with humans then birds and fish, had developed the ability to abstract from (local, low-level) visual cues to (global, higher level) letter and bigram *representations* (Grainger et al., 2012) .

The results of a follow-up study show that baboons are also sensitive to the transposed letter effect, tending to categorize non-words created by the transposition of the middle letters as words, while correctly classifying control targets created by letter substitution as non-words. In humans, sensitivity to the spatial organization of letters in visual word recognition is typically attributed to flexible orthographic coding (abstraction). This is often considered a characteristically human capacity, such that the performance of the baboons here is particularly noteworthy (Ziegler et al., 2013).

However, pigeons in this paradigm show the same sensitivity to the information structure of words as humans and baboons. Although the pigeons master fewer words than the baboons, they too were able to discriminate actual words from non-words, and perhaps most strikingly, pigeons are also sensitive to the transposed letter effect, indicating that avian brains are as capable of learning to read as the brains of baboons. Notably, since pigeons brains lack a ventral pathway, this suggests that this structure itself may not be a necessary prerequisite for orthographic learning (Scarf et al., 2016).

Finally, simulations have been used to study the mechanisms required to produce these results. Using a simple discriminative learning model Linke et al. (2017) examined whether the behavior of the baboons in these experiments could be acquired from training on low-level gradient cues. The model successfully replicated their performance on all of the reported measures, including the transposed letter effect. Taken together, the results of the pigeon and simulation studies are particularly interesting, because results from visual search experiments suggest that while humans exhibit a global bias (categories), animals tend to exhibit a local bias (features), albeit there is evidence that contingent on their experience, at least some species of animals might exhibit global preferences (Avargues-Weber et al., 2015))

All of which raises the question of what we are to conclude about reading from comparative studies? The answer, as ever, seems to be 'it's complicated:' on one hand, we might concluded that baboons, pigeons and even models learning from low-level features appear to be capable of forming the kind of global, high-level representations thought to be required for reading (which seemingly contradicts many previous findings that suggest that animals do not learn these kind of strategies in visual search); on the other hand, since it appears that both pigeons and a low level, bottom-up learning model are capable of learning representations sufficient

to succeed at these tasks, we might conclude that reading doesn't require global, high-level representations after all. In truth, because even pigeons and simple learning model form abstractions, on the basis of these results alone one can only guess which of these is right.

In what follows, we adopt a different approach to explaining these results. Whereas much of the research summarized above focused on the representations formed by learners (be they human or baboon), in some ways, this question is ill-posed. In this (and many other cases where we seek to explain learning effects) the important question is not, 'what representations are formed?' but rather what is the structure in the environment that gives rise to those representations? In other words, the important question here is not "are baboons and pigeons capable of forming human-like representations?" but rather: does the performance of the baboons and pigeons in fact reflect the complex structure of the letter sequences in human orthographic codes, which are likely to have culturally evolved to maximize the discriminability and learnability of the information provided by words in context?

To answer this question, we will examine the structure of the environment from which the visual targets used in comparative reading experiments are drawn. In particular, we will analyze the degree to which the fact that the words and non-words used in these studies were drawn from a highly structured environment (a human orthographic code) has imbued them with hitherto unnoticed latent structure. In combination with the results of a simulation experiment, we shall examine whether the environmental structure we identify can account for the apparent contradictions in previous findings.

7.3 The Empirical Structure of Communicative Distributions

In recent years our understanding of the structure of the linguistic environment has changed considerably, in large part due to the development of massive speech and text corpora, and the tools to analyze and mine them, shifting mainstream computational approaches to language from a concentration on logic (and considerations of the poverty of the stimulus) to a concentration on learning (and an appreciation of the richness of the stimulus) in just two decades. Such has the impact of these methods been that Liberman (Zimmer, 2012) has likened their effect on research in linguistics to that of the invention of the telescope in physical sciences.

One recent finding that is particular relevant for current purposes is the amount of local detail that linguistic information structures encode. It has long been known that

human communicative distributions are far from uniform (Estoup, 1916; Zipf, 1949), such that cumulative lexical distributions have a 'Zipfean', power-law distribution. However, recent analyses show that Zipfean distributions merely reflect the effect of aggregation over functional communicative distributions, with closer analysis revealing that the distributions that language users actually encounter in context are geometric across multiple levels of linguistic description (Ramscar, 2019; Linke and Ramscar, 2020a).

From a communicative perspective, these distributions indicate that the 'stimulus' that language learners are exposed to is far from impoverished. Instead, they show that human communicative codes comprise complex aggregates of hierarchically organized (nested) distributions that seem to be optimized for learning and transmission (geometric distributions are sampling invariant, such that they support the learning of convergent probabilistic models even in learners with vastly different levels of experience, supporting the acquisition of joint-expectations across a community of learners). This pattern of distribution is found at all levels of description, from argument structures (Ramscar, 2019) to word initial contrasts in conversational speech (Linke and Ramscar, 2020a).

By contrast, from what we understand about the communicative codes of animals, their signals do not usually exhibit the same kind of hierarchical structure in their distributions. Specifically, when the frequency distributions of animal signals are aggregated they resemble the local distributions discriminated by context in human codes (Hailman et al., 1985; McCOWAN et al., 1999), as opposed to the distributions that emerge in human codes when distributions are aggregate across the hierarchy of communicative contexts.

7.3.1 What is learning?

Having established some characteristic of the kinds of codes humans and animals learn, we next turn to how they learn them. Humans share their basic learning mechanisms with other animals, allowing animal models to provide insight into human learning (Ramscar et al., 2013c). These studies suggest that the learning mechanisms seen in human and animals are error-driven, a mechanism that has been subject to a huge amount of computational research that provides many insights into the capabilities of and constraints on this kind of learning in real-world situations (LeCun et al., 2015). From this perspective, the processes by which humans and animals learn about the world is best characterized in terms of expectation and uncertainty reduction. Experience serves to organize and manage a learner's expectations about the world in context, serving to reduce learners' uncertainty about their environments over time.

Although there is considerable evidence that human communicative learning relies on the mechanisms we share with other animals, some specific aspects of human communicative learning are worth considering in relation to the use of animal models in the study of language. First, while animal and human learning might rely on similar mechanisms, the protracted pattern of development of the human learning architecture is markedly different to that of other animals, and in particular, its limiting of the behavioral flexibility of early learners appears to be specifically adapted to the acquisition of the conventions that constrain and guide human communication (Ramscar and Gitcho, 2007). Second, linguistic information structures have evolved in a cultural environment constrained (and in part shaped) by the characteristics of immature learners. Third, unlike baboons and pigeons human learners are explicitly taught to read, a process that involves more than merely discriminating words from non-words – for human learners, the important task is learning to discriminate words from words, because for human readers words serve a functional, communicative purpose. Fourth, and perhaps most importantly, humans learn to read words in context, and it is in context that words actually achieve their functional purposes.

Thus while baboons and pigeons clearly 'learn words', these considerations raise questions about the degree to which what they learn is the same as humans. Does baboon (and pigeon) performance in reading experiments show they learn what human readers learn? Or is it the case that baboons and pigeons are simply sensitive the structure of letter sequences and orthographic codes, sequences that have otherwise evolved to maximize the ability of human learners to discriminate and learn communicative signals in context?

To address these questions, we first reconstructed a simulation of learning in this task previously reported in Linke et al. (2017). In contrast to this earlier work, however, our goal here was not to simulate performance, but rather to use the model as a means to explore the information available to (at least baboon and pigeon) learners in this task. Accordingly, we first ran the simulation, and then conducted a series of analyses that sought to relate what the model learned and how it captured baboon and pigeon behavior to the environmental information that actually shaped the behavior of the model.

7.4 Simulation Study: Learning from Visual Cue Distributions in English 4-letter Sequences

In our simulation of the baboon reading experiments representations of the training words (described in detail below) served as inputs to a two-layer network, which learned to classify them into words and non-words using the simplified Delta rule

(Widrow and Hoff, 1960; Rescorla, 1972; Stone, 1986). To simulate learning at the individual baboon level, a separate network with word and non-word output nodes was constructed for each baboon using the the training sequences employed in the actual study.

Test words were constructed following the procedure in Ziegler et al. (2013): non-words were created by randomly selecting words from the original training set (Grainger et al., 2012) and then transposing or substituting (exchanging a vowel for a vowel or a consonant for a consonant) center letters.

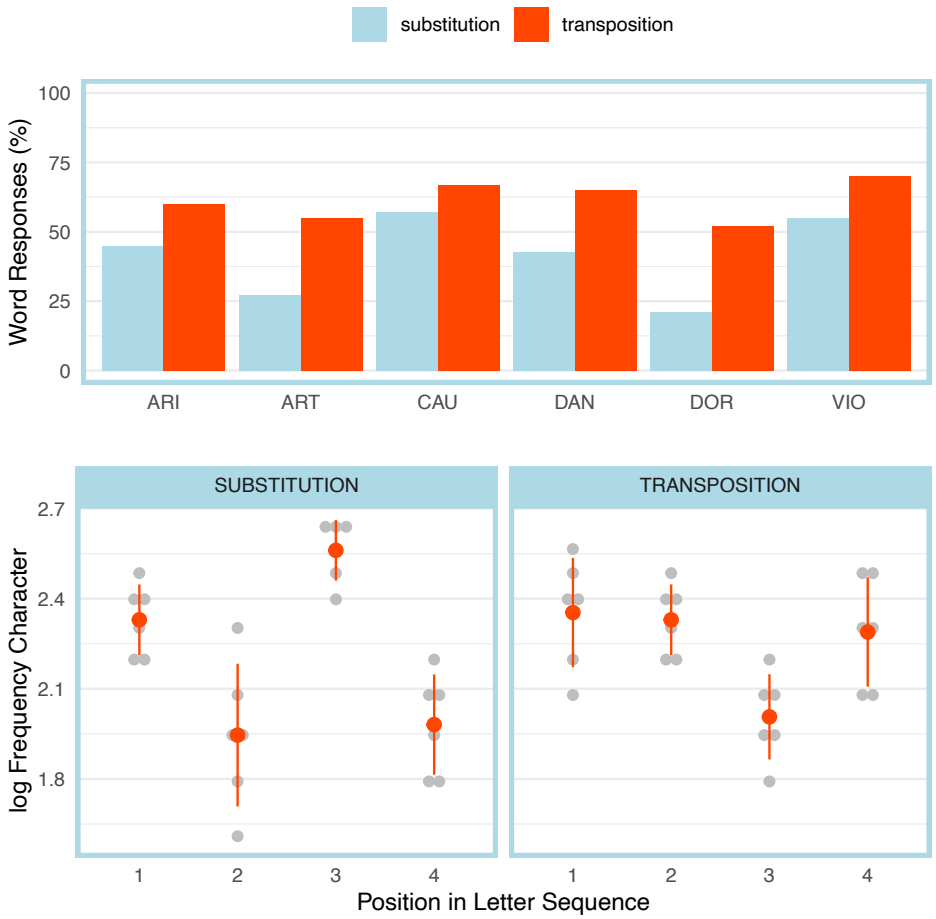

Fig. 7.1: **Top:** Simulation results for transposed letter condition (blue) and letter substitution (orange) replicate the experimental results reported by Ziegler et al. (2013) with nonwords derived through letter substitution being classified as words more often, **bottom:** the corresponding letter distributions for transposed (right) and substituted (left) condition for each of the 6 subject models show that letter substitution leads to less variability in the distribution of uncertainty (log transformed frequency of distinct letters) across word positions. Theoretically, the animal (or model) can learn to classify the presented string by estimating the likelihood of seeing novel/unexpected combinations of visual features within the string.

7.4.1 Oriented Gradients as Cues to Learning

Words in the test and training sets were represented using discrete gradient orientation descriptors of the low-level features of the printed words when treated as individual images. The features encode gradient magnitude, gradient orientation, and spatial location - the algorithm (Dalal and Triggs, 2005) computes magnitude of gradient orientation in densely distributed locations of an image. For this, each image is divided in a 10 by 4 grid containing non-overlapping cells of fixed size. For each cell, we extract the gradient magnitude at each pixel contributing a weighted vote to 9 gradient orientation bins of 20 degrees each, adding a weighted contribution to the four neighboring orientation bins and the respective bins of four neighboring cells.

The resulting histograms values were normalized (with the regularized L2-norm) across cells, yielding for each image a gradient-based vector with 40 $\times$ 9 = 360 values in $[0, 1]$. To capture locality and orientation, we add to each value a head argument encoding the feature descriptor index, representing the cell position and gradient bin. The 8139 distinct word and non-word targets of the baboon study generated a total of 14,476 unique local gradient orientation features.

7.4.2 Simulation results

Each network was trained on the sequence of words in the order presented to that baboon in the experiment and the performance of the networks was tested. The simulations successfully reproduced the effects reported for the baboons: non-word strings obtained by the transposition of two middle characters elicit more word responses ($mean = 66.15$, $sd = 7.77$) than non-words created by letter substitution ($mean = 43.41$, $sd = 12.57$).

7.4.3 What can we learn from from this?

Since all cognitive modeling ultimately relies on analogy, from a theoretical perspective, is seems clear that examining what we can learn from the factors that determined what the model learned, and how these factors shaped the models' *behavior* can provide far more insight into the factors that actually shape reading performance (in both humans and baboons) than considering whether the model captured (by analogy) the performance of the baboons (which it did).

To begin this process, we will compare the cumulative frequency distributions of low-level cues, letters and letter bigrams to the distributions of cues at the positions

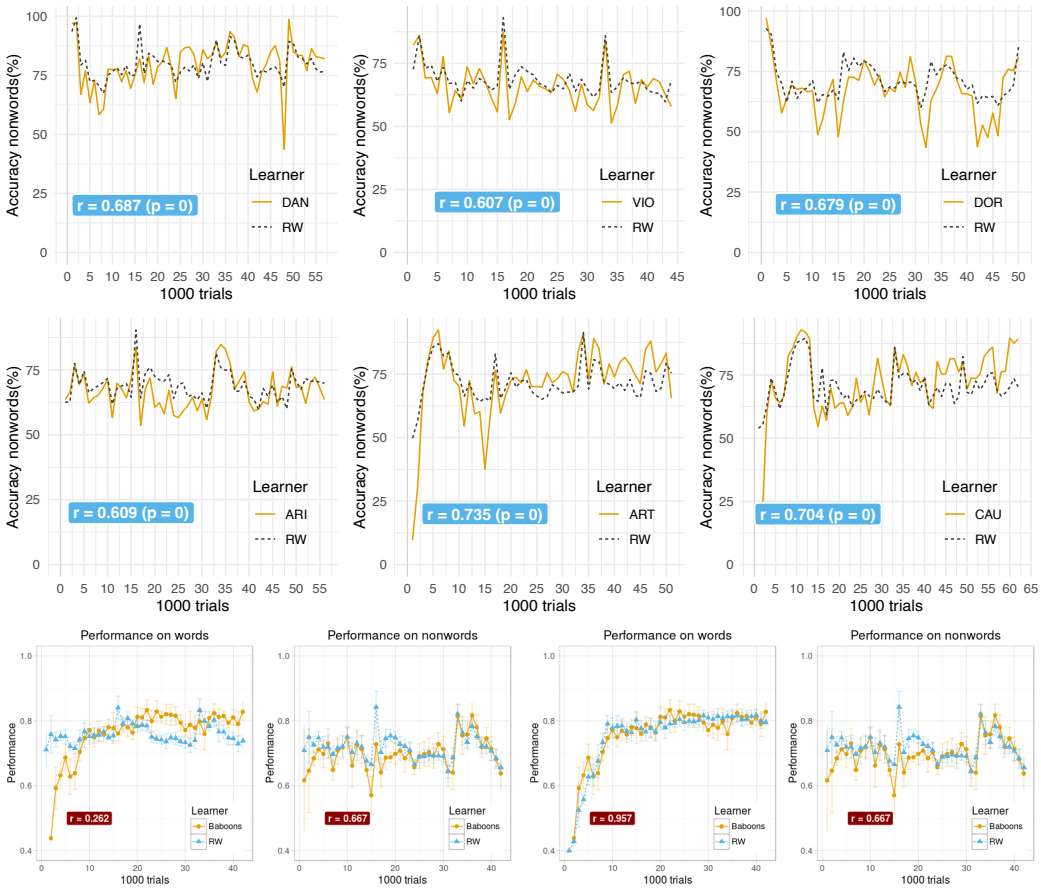

Fig. 7.2: Top rows: Block-wise performance of individual baboons (yellow) on non-word stimuli plotted against the performance of a network trained on the same set of stimuli (gray), **Bottom rows:** Block-wise word accuracy for all baboons and two different baboon models. The model on the left categorizes the stimulus based on activation weights (evidence), the model on the right side of the plot is allowed to 'guess randomly' under uncertainty (when the differences between word and non-word activations are small) and learn from its own behavior subsequently. The model performance on non-words does not change, the guessing model that creates a copy on its own (noisy) behavior performs more baboon-like on word stimuli.

they were presented at during the training experiment. Specifically, we will examine how the shape of the distribution and the slope of the distribution vary with position for words and non-words, respectively. We will then examine the spatial information provided by the gradient features in training sets, again comparing words and non-words. In particular, we will compare the information in the distribution of low-level and high-level cues (letters and letter bigrams) in order to establish their influence on the model's performance in relation to the different kinds of test words. Finally we will explore whether the differences revealed by these analyses can explain subsequent performance on the test items.

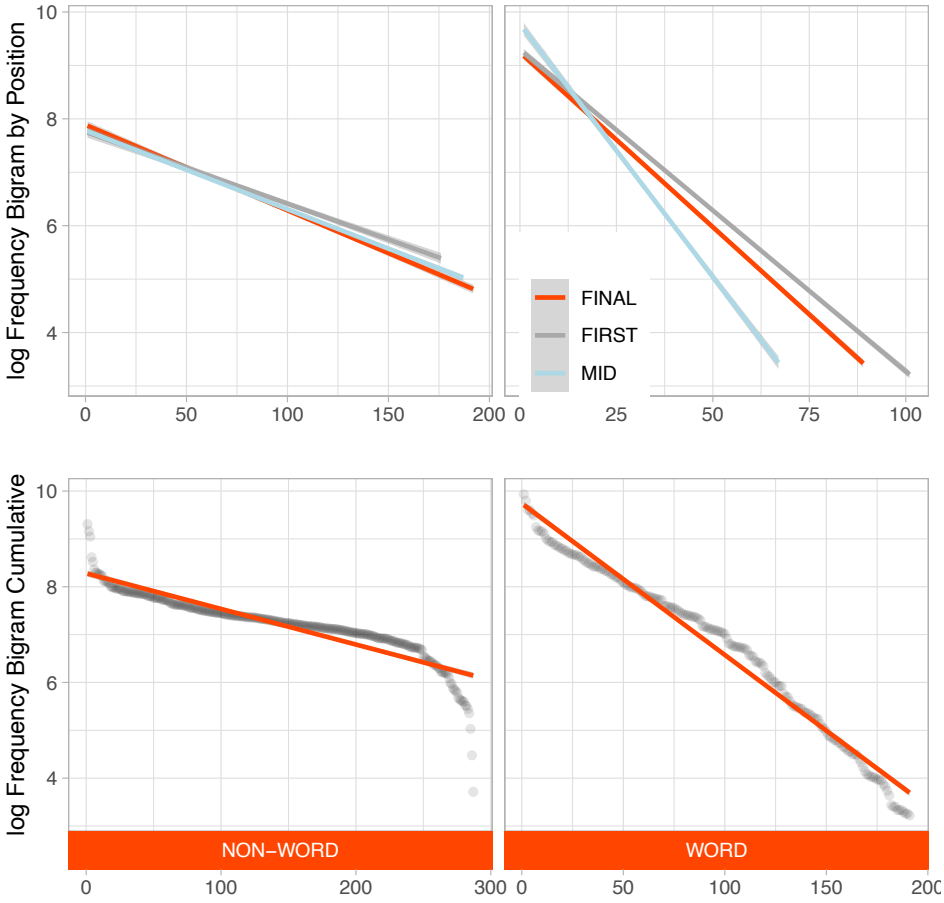

Fig. 7.3: The log frequency-rank distribution of character bigrams by position they take in the word (top) and the cumulative distribution of bigrams (bottom) for words (right) and non-words (left). Considering bigram distributions in their spatial context increases the contrast (differences in the distribution of information, reflected in slope differences) between bigrams in words in a way it does not in non-words. This means that in random letter strings (non-words) the information (and uncertainty) is distributed equally across the positions, while the uncertainty distribution of substrings in words is structured by usage to provide more information at the word boundaries. The cumulative distribution of word bigrams appears to be an aggregate of bigrams from structurally distinct contrast clusters (hence the differences in slopes).

7.4.4 Distributional Analysis of Training Data

As we noted above, functional distributions in human communicative codes (the sets of forms that occur in context at various levels of description) are geometric. These distributions serve to optimize both the alignment of expectations among users of a code, and the efficiency of signaling. Notably, similar distributions have been observed in the frequency distributions of animal signals (Hailman et al., 1985; McCOWAN et al., 1999), suggesting that animal codes support similar functions. Thus although the prior experiences of human and baboon readers will vary dra-

matically, it seems reasonable to assume that both will share a bias for learning information structured in this way.

Tab. 7.1: Frequency Distribution of Letter Bigrams.

Words	R^{2*}	Slope	Entropy	Perplexity
Initial	0.988	-0.06	4.588	24.051
Mid	0.982	-0.095	4.163	17.914
Final	0.988	-0.065	4.389	20.952
Cumulative	0.851	-0.004	4.601	24.268
Non-Words	R^{2*}	Slope	Entropy	Perplexity
Initial	0.808	-0.013	5.073	33.661
Mid	0.946	-0.015	5.157	35.679
Final	0.904	-0.016	5.120	34.776
Cumulative	0.935	-0.009	5.250	38.055
*fit geometric				

Accordingly, our first analysis sought to establish the degree to which the stimuli used in training in these experiments actually conformed to these expectations. To do this, we analyzed the frequency distributions of letter bigrams by position they take in the letter sequence, distinguishing between word initial, center and word final bigram and compare this to cumulative bigram distribution for word and non-word targets. There are 101 unique letter bigrams in words and 192 in the non-words presented to the animals.

These differences reflect the constraints imposed by the combinatoric nature of human communicative systems - bigrams are not random letter combinations where new words are scrambled together whenever a new word is required - they appear to have evolved (culturally) in response to the pressures that shape the structure of codes, e.g., the discriminability of forms in context, learnability etc. Analyses have shown that not only are the forms discriminated by context distributed systematically, but also that the patterns of lexical co-occurrence that provide that context itself share that same systematic structure. These analyses indicate that information (and thus uncertainty) is systematically structured across contexts in these codes, and that this kind of 'nested' structure can be observed across many levels of description/analysis (Ramscar, 2019; Linke and Ramscar, 2020a). Since human communicative codes are combinatoric, the contribution of any given distribution (of phones, forms, arguments, n-grams etc.) also depends on its role in the larger system (which will be reflected in the amount of uncertainty associated with the contexts in which a given distribution of items occurs).

Consistent with this, the current analysis revealed that the distributions of the bigrams at all three positions (initial, center, final) in the target words closely approximate a

geometric distribution (Table 7.1), but with different slopes that reflect the uncertainty associated with word boundaries (i.e., the information they contribute in reading) as compared to the middle of words. Consistent with previous observations of the effects of aggregating functional distributions, the cumulative distribution (all word bigrams, independent of their position) fit deviated from the geometric fits seen above.

These patterns reflect the combinatoric constraints set by English orthography on the distribution of bigrams, and in particular, the need for sublexical contrast be distributed across sequences, and the information conveyed by word boundaries (these in turn are defined by spaces, which play a critical organizing role in the English writing system).

By contrast, all of the bigram frequency distributions of non-words (initial, mid, final, and cumulative) deviate from the empirical optimum (Table 7.1). In other words, the information provided by bigrams differs between word and non-words, and critically, these differences are particularly notable at the boundaries of words (i.e., where the information provided by words combines with that provided by spaces) . As a consequence, words and non-words not only differ in their discriminability, they also differ with regards the spatial distribution of the information that contributes to their discriminability. Real words distribute this information (measured by perplexity in Table 7.1) differently across bigrams depending on their location within a word, however these informative differences are largely absent in the non-words. Since what drives learning is the information in the training items, these differences indicate that as compared to words, the topology defined by bigrams in non-words will become increasingly uninformative over time.

7.4.5 Distributions of low-level features

Analysis of the gradient feature distributions revealed geometric distributions in both word and non-words. Further analysis of the low-level visual features by position within the word also revealed no differences: word and non-word features closely approximating a geometric fit across the four letter positions ($R^2 > 0.97$).

However, the far larger number of features in relation to the far larger set of non-words employed in the experiments results in their distributions having shallower slopes, which reveal that on average, non-word features convey less discriminatory information than word features. The by-position differences observed in bigrams are present in both words and non-word features and follow the same trends (at this low level, there is more information at the word boundaries in both words and non-words).

We do, however, find marked differences in the way low-level features are distributed across the representations of the training words (i.e. *differences in topology* in the distribution of grid cells used to represent the words, Figure 2). Specifically, the distribution of these low-level cues is far more irregular across the non-words than it is across the words, such that there are larger differences in the discriminatory information provided by individual cells in non-words as compared to words. This means that the spatial distribution of information provided by low-level cues across the low-level representations of the training words provides a great deal of information about the fact that non-words are not words.

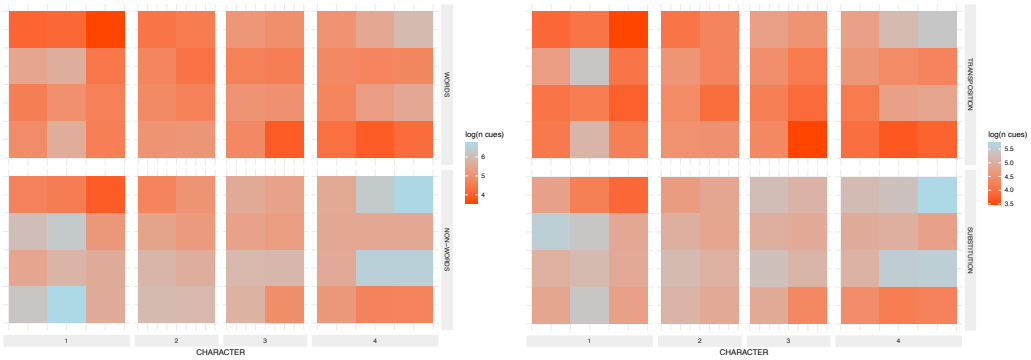

Fig. 7.4: Left: Distribution of gradient information across individual grid cells by letter position: words (top) provide more contrast at the word boundaries than non-words (bottom), also information density is higher in words. **Right:** Contrast in information between targets from the simulation: non-words derived by letter substitution (bottom) provide less information at the word boundaries, while non-words derived by letter transposition (top) maintain a word-like pattern of visual contrast distribution.

Consistent with this analysis, an examination of the blockwise performance of the models trained on low-level features in the simulation and the actual performance of the baboons revealed the correlations between model and animal behavior to be far higher for non-words ($mean = 0.670$, $sd = 0.05$) than for words ($mean = 0.479$, $sd = 0.180$), which indicates that to some extent, word versus non-word discrimination could be learned from these low-level features alone.

On the other hand, the analysis of spatial distributions of low-level features reveals that non-words created by letter-transposition Figure 3 show a distribution pattern that is almost identical to words in Figure 2 (indicating this discrimination could *not* be learned from low-level features) while targets created by letter substitution closely resemble the pattern observed in non-words (indicating this discrimination could).

In other words, whether words and non-words are discriminable (and the level of representation required to discriminate them) may depend on the actual non-word discrimination task (transposition versus substitution) and the specific characteristics of the non-words employed in them.

7.5 General Discussion

We analyzed the distributional structure of the training and test sets presented to baboons in a reading experiment. These analyses revealed that the information available in the words and non-words tested differs, and in particular, that there were important differences in the spatial location of information in words and non-words. Moreover, depending on the way words were transformed into non-words, these differences manifest themselves at different levels of representation. Since these transformations defined different 'reading effects' this means that baboons' performance on different reading tasks appears to be best accounted for by features or representations learned at different levels.

All of the above suggests that understanding reading and the skills associated with human communication may require a more nuanced approach than that of asking simple nature versus nurture questions. That is, the critical question may turn out to be not *are the representations associated with reading learned or innate*? but rather, is there something about the socially evolved structure of the linguistic input that makes reading (and also more general communicative skills) possible?

All of which raises questions: what does it mean to say baboons exhibit human-like behavior; what we can infer about human processes and representations from this? Some obvious answers are: baboon (and human) behavior in these experiments results from their exposure to stimuli that have been shaped by human cultural evolution, which they learn about via similar mechanisms. Yet the contexts in which baboons and humans 'learn to read' (and their prior learning in doing this) differ massively: it is highly unlikely they both learn 'the same' representations. This suggests that using animal (and human) models to study language requires a more nuanced approach. Rather than asking *which species can learn what,* it might be more useful to focus on establishing what it is about different cultures and patterns of development that allows species to learn to communicate as they do.

Conclusion: It's all in the code

8

> *Further conceive I beg, that a stone, while continuing in motion, should be capable of thinking and knowing, that it is endeavoring, as far as it can, to continue to move. Such a stone, being conscious merely of its own endeavor and not at all indifferent, would believe itself to be completely free, and would think that it continued in motion solely because of its own wish. This is that human freedom, which all boast that they possess, and which consists solely in the fact that men are conscious of their own desire, but are ignorant of the causes whereby that desire has been determined.*
>
> — **Baruch Spinoza**

This work presents findings which indicate that the information content of speech signals varies with uncertainty and speakers' expectations. We show how both collective and individual 'computations' contribute to uncertainty reduction and efficient transmission of information in human communicative codes. Communicative conventions seem to be maintained by distributed, collective processes that ensure the predictability of shared codes.

The results of these analyses suggest that far from being articulatory noise, form variants optimize the distribution of speech contrast to maximize the communicative efficiency of speech in context. That is, despite what might seem to be the disorderly nature of conversational speech and the large differences in the way individual speakers articulate phrases, words, and speech sounds (i.e. speech forms), the aggregate distributions of sequence lengths and sequence positions, syntactic forms and phonetic segment clusters they discriminate between, converge on exponential distributions. This is remarkable, because **??**exponential distributions satisfy two important functional requirements of communication.

First, they provide the kind of structure to support initial signal organization and facilitate learning. Second, they are memoryless, which means that speakers

learning from this kind of distribution will acquire and maintain sufficiently similar expectations. Notably, the highly conventionalized and stable aspects of linguistic structure, such as utterance and word length, parts of speech, function words, and phonetic-acoustic inventories aggregate to globally informative memoryless distributions. That is, despite local fluctuations in the extent to which individual speakers conform to shared conventions, collective performance approaches the theoretical optimum. In addition to that, the distributed lexical clusters these surface-level regularities discriminate between also approximate exponential distributions. In other words, regularities at various levels of description appear to systematically organize the system of speech contrasts into a memoryless distribution of structured memoryless distributions. What is memorylessness, and why is this important?

As we have noted at various points in this thesis, memorylessness describes a formal property of certain event distributions where differences in knowledge about the events that have occurred prior to a certain point in time confer no advantage. When a distribution of events is memoryless, it follows that despite differences between local samples speakers learn from, global (shared) event probabilities will be unaffected by any individual knowledge of the history of the process (i.e. local bias). This property suggests that despite the variability in individual speakers' experience and the differences in the relationships between individual words and the structure of articulated sequences that arise as a consequence of this variability, speakers somehow manage to collectively maintain a sufficiently stable distribution of conventionalized patterns of communication (forms). This enables them to acquire sufficiently similar expectations and communicate efficiently. The critical aspect of this finding is that communicative conventions rest on stable, systematic patterns of co-occurrence between words, syllables, and speech sounds. This implies that the variability in the way individual speakers (mis)articulate words, syllables, and speech sounds is itself systematic. This is puzzling, how does this collective systematicity emerge out of what seems to be noise?

Further findings suggest that the key to systematic variation lies in the structural differences in the distribution of types from different lexical categories. Initial analyses (presented in chapter 3) show that speech seems to be subcategorized by regular patterns of co-occurrence at multiple levels of abstraction: grammatical, lexical, and sublexical. The extent to which different lexical categories (e.g. verbs and nouns) sub-categorize at different levels interacts with the lexical productivity of the category (the type/token ratio) and the variance in the recurrence rates of the types from this category. Verbs, which are less lexically productive than nouns, tend to have more stable recurrence rates (Altmann et al., 2009; Genzel and Charniak, 2003). They also tend to subcategorize in more stable, grammatical clusters (verb alternation classes, time and tense, etc. (see Levin, 1995; Ramscar, 2019)). Because usage patterns behind grammatical conventions tend to not change substantially within the average

lifespan, language users knowledge about them will not increase substantially over time. The constraints that grammar imposes on verbs suggest that the uncertainty at the transition between the preceding argument of the argument frame and the verb will, on average, decrease with experience (or the size of the observed sample). By contrast, the variance in the conditional probabilities of nouns and argument frames they appear in increases as the speaker experience grows. Nouns become increasingly diversified by the lexical context and tend to sub-categorize in semantic clusters that are less stable in time and, therefore, across individuals.

The structural differences mentioned above also cumulatively increase the differences between the distributions of English verbs and nouns. In English verbs, the head of the distribution, which represents very frequent verbs, grows faster than the rest of the distribution as the sample size (observation time) increases. In nouns, the opposite is the case: the growth in the high-frequency head of the noun distribution is relatively stable, while the number of unique nouns in the low-frequency tail of the distribution increases disproportionately fast with the sample size. That is, verb and noun distributions, when considered independently, are not invariant under scaling (they do not follow power laws). Verb and noun aggregates, by contrast, approach power laws (which are defined by scale invariance). Scale invariance implies that when all words are considered together, both frequent and rare words become very frequent and very rare in relation to the rest of the distribution at approximately the same pace as the sample size increases. This is not the case when words from different categories are considered independently. Why is this is important?

The functional forces behind scale-free distributions bear broad implications for the structure and dynamics of complex systems (including systems of communicative contrasts). Provided that human communicative networks operate by principles that apply to other aspects of human behavior, the dynamics behind the structure and the distribution of human communications can help illuminate processes that shape other more or less scale-free distributions. Moreover, as the maintenance of structure in complex systems appears to be shaped or at least correlated to the principles of hierarchical reorganization in living systems (see Flack et al., 2013), the organizing principles of communicative distributions may help reveal general principles that allow systems to reorganize at multiple levels without losing their scaling properties. For example, explaining the conditions under which the tails of communicative distributions grow at a similar pace (and the conditions under which they do not) may tell us how scaling and symmetry in complex systems are maintained. The distributional learning approach that serves as a framework for this thesis predicts that the distribution of verbs and nouns in human speech sequences is a response to two competing sources of functional pressures: the maintenance of a sufficiently invariant structure and mutual predictability on one side, and lexical innovation and discriminability on the other. Mutual predictability is maintained

through stable patterns of co-occurrence between word forms from different classes. These collocational regularities govern semantic and syntactic subcategorization. Discriminability can be maintained through reconfiguration of existing patterns of co-occurrence – by using familiar words in unfamiliar contexts and modifying the word order – or by introducing new word forms and adapting articulations.

Further analyses reveal that these quantifiable differences in the structure of lexical subcategories systematically shape sublexical variation in speech signals. First, the analyses presented in chapter 4 indicate that variation in the way words are articulated is a consequence of the diversity of lexical contexts a word appears in and not, as previously suggested, word probability in context (Bell et al., 2009; Piantadosi et al., 2011; Gibson et al., 2019; Levy, 2008; Hall et al., 2018). This indicates that form variability is a response to local contextual uncertainty (knowledge about the distribution of transitional probabilities at a certain point in time/sequence) and is not a global, systemic effect (relying on knowledge about transitional probabilities in the context of the aggregation of all distributions). Further, we presented evidence that form deviation (phoneme and morpheme exchanges) leads to formation of nearly identical memoryless speech segment distributions across grammatical contexts that differ enormously when quantified in terms of token counts and numbers of individual types. Specifically, the distributions of initial segments in lexical categories, like common nouns and pronouns, that host 3752 vs 31 unique word types, respectively, are nearly identical. How does a system that is distributed across individuals acquire such a structure?

In a further step, we show that speech segment deviation develops predictably as speakers get older. The probability of form deviation changes as a function of sequence position and the duration of the preceding pause. These two relatively stable quantities (sequence and pause) show a differential impact of lifelong experience in verb and noun production. In the more lexically productive nouns, experience changes both where and how much articulations vary with contextual uncertainty. Specifically, speaker experience seems to strengthen patterns of articulation when the uncertainty is low and dissipate patterns of articulation under high uncertainty. In the relatively stable verbs, by contrast, the change in patterns of articulation is a matter of degree. Variation decreases with contextual uncertainty – the pattern itself does not change, only the magnitude and the resolution do. Another implication of these analyses presented in chapter 6 is that experience increases the sensitivity to information and uncertainty, such that the effect magnitude peaks at the more informative word boundaries. These findings suggest that articulations and, with them, speech forms develop continuously across the lifespan and that variation in the way words are pronounced reflects the contextual uncertainty inherent to individual speakers' experience of the linguistic code.

This led to a new, more focused question. The fact that experience is distributed in space and time, yet seemingly random variation in individuals is systematic when considered together, indicates that the signals and the uncertainties their transmission entails must be structured systematically in relation to some invariant quantity. If these predictable patterns of form deviation arise from systematic developments in the rate at which changes in speech signals are discriminated, which part of the signal maintains alignment in rates?

To explore this question, I propose pause duration as an objective function of time and contrast its lifespan development against sequence position. I find that pause distribution in speech in Korean and English are both identical and critically memoryless. This last property suggests pauses can provide structure and, at least in speech, allow us to share our experience of time. Crucially, the pauses offer a structural template that is independent of individual experience. Their distribution enables speakers whose experience differs widely to nevertheless build and maintain sufficiently stable expectations about the occurrence of relevant changes in the speech signal. Taken together, this suggests that linguistic codes are bounded by the interaction between two different sources of information – the memoryless distribution of pauses and the structure of articulations that are shaped by the collective experience.

These findings raise many questions regarding some widely held linguistic assumptions. For decades, in the absence of any way of meaningfully collecting or analyzing large sample of speech data, linguistic theories have simply assumed the unequivocal existence of speech units such as words and speech sounds. This has led to the development of many formal theories of language based upon them. By contrast, the results presented here suggest that language emerges out of noisy signals that are structured by the coordination of shared speaker expectations in time. They indicate that the efficiency and productivity of human communication systems emerge from coordination and cooperation between speakers. This, in turn, appears to undermine the assumption that languages are formal systems. Instead, the results presented in this thesis support a very different conclusion: that even in their structures, languages are social systems that inevitably develop and change according to the needs of communities of speakers.

Bibliography

Ackermann, Hermann, Klaus Mathiak, and Axel Riecker (2007). „The contribution of the cerebellum to speech production and speech perception: clinical and functional imaging data“. In: *The cerebellum* 6.3, pp. 202–213 (cit. on p. 73).

Altmann, Eduardo G, Janet B Pierrehumbert, and Adilson E Motter (2009). „Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words“. In: *PLOS one* 4.11 (cit. on pp. 81, 158).

Amitay, Sygal, David JC Hawkey, and David R Moore (2005). „Auditory frequency discrimination learning is affected by stimulus variability“. In: *Perception & psychophysics* 67.4, pp. 691–698 (cit. on pp. 16, 28).

Amitay, Sygal, Amy Irwin, and David R Moore (2006). „Discrimination learning induced by training with identical stimuli“. In: *Nature neuroscience* 9.11, pp. 1446–1448 (cit. on pp. 16, 29).

Andruski, Jean E, Sheila E Blumstein, and Martha Burton (1994). „The effect of subphonetic differences on lexical access“. In: *Cognition* 52.3, pp. 163–187 (cit. on p. 113).

Arbesman, Samuel, Steven H Strogatz, and Michael S Vitevitch (2010). „The structure of phonological networks across multiple languages“. In: *International Journal of Bifurcation and Chaos* 20.03, pp. 679–685 (cit. on pp. 16, 43).

Arnold, Jennifer E, Maria Fagnano, and Michael K Tanenhaus (2003). „Disfluencies signal theee, um, new information“. In: *Journal of psycholinguistic research* 32.1, pp. 25–36 (cit. on pp. 3, 39, 113).

Arnold, Jennifer E, Carla L Hudson Kam, and Michael K Tanenhaus (2007). „If you say thee uh you are describing something hard: the on-line attribution of disfluency during reference comprehension.“ In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33.5, p. 914 (cit. on pp. 39, 58, 113).

Arnon, Inbal and Uriel Cohen Priva (2014). „Time and again: The changing effect of word and multiword frequency on phonetic duration for highly frequent sequences“. In: *The Mental Lexicon* 9.3, pp. 377–400 (cit. on p. 59).

Arnon, Inbal and Michael Ramscar (2012). „Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned“. In: *Cognition* 122.3, pp. 292–305 (cit. on pp. 32, 35, 104).

Avargues-Weber, Aurore, Adrian G Dyer, Noha Ferrah, and Martin Giurfa (2015). „The forest or the trees: preference for global over local image processing is reversed by prior experience in honeybees“. In: *Proceedings of the Royal Society B: Biological Sciences* 282.1799, p. 20142384 (cit. on p. 144).

Aylett, Matthew and Alice Turk (2004). „The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech“. In: *Language and speech* 47.1, pp. 31–56 (cit. on pp. 1, 3, 14, 19, 70).

Baayen, Harald and Maja Linke (2020). „An introduction to the generalized additive model“. In: *A practical handbook of corpus linguistics*. Springer (cit. on pp. 120, 129).

Baayen, R Harald (2001). *Word frequency distributions*. Vol. 18. Springer Science & Business Media (cit. on p. 47).

Babel, Molly (2010). „Dialect divergence and convergence in New Zealand English“. In: *Language in Society* 39.4, pp. 437–456 (cit. on p. 16).

Baese-Berk, Melissa M, Laura C Dilley, Molly J Henry, Louis Vinke, and Elina Banzina (2019). „Not just a function of function words: Distal speech rate influences perception of prosodically weak syllables“. In: *Attention, Perception, & Psychophysics* 81.2, pp. 571–589 (cit. on pp. 6, 72, 77, 105).

Baese-Berk, Melissa M and Arthur G Samuel (2022). „Just give it time: Differential effects of disruption and delay on perceptual learning“. In: *Attention, Perception, & Psychophysics* 84.3, pp. 960–980 (cit. on p. 76).

Bard, Ellen Gurman and Anne H Anderson (1983). „The unintelligibility of speech to children“. In: *Journal of Child Language* 10.2, pp. 265–292 (cit. on pp. 71, 74, 104).

Bell, Alan, Jason M Brenier, Michelle Gregory, Cynthia Girand, and Dan Jurafsky (2009). „Predictability effects on durations of content and function words in conversational English“. In: *Journal of Memory and Language* 60.1, pp. 92–111 (cit. on pp. 1, 3, 38–40, 59, 70, 160).

Bell, Alan, Daniel Jurafsky, Eric Fosler-Lussier, et al. (2003). „Effects of disfluencies, predictability, and utterance position on word form variation in English conversation“. In: *The Journal of the acoustical society of America* 113.2, pp. 1001–1024 (cit. on pp. 1, 3, 14, 19, 58, 70, 113).

Bentz, Christian, Douwe Kiela, Feli Hill, and Paula Buttery (2014). „Zipf's law and the grammar of languages: A quantitative study of Old and Modern English parallel texts“. In: *Corpus Linguistics and Linguistic Theory* 10.2, pp. 175–211 (cit. on pp. 48, 49).

Best, Catherine C and Gerald W McRoberts (2003). „Infant perception of non-native consonant contrasts that adults assimilate in different ways“. In: *Language and speech* 46.2-3, pp. 183–216 (cit. on p. 103).

Blei, David M, Andrew Y Ng, and Michael I Jordan (2003). „Latent dirichlet allocation“. In: *Journal of machine Learning research* 3.Jan, pp. 993–1022 (cit. on p. 14).

Blevins, James P, Farrell Ackerman, Robert Malouf, and Michael Ramscar (2016). „Morphology as an adaptive discriminative system“. In: *Morphological metatheory*, pp. 271–302 (cit. on pp. 6, 72, 110, 116, 191).

Blevins, James P, Petar Milin, and Michael Ramscar (2017). „The Zipfian paradigm cell filling problem“. In: *Perspectives on morphological organization*. Brill, pp. 139–158 (cit. on pp. 35, 116, 191).

Bóna, Judit (2011). „Disfluencies in the spontaneous speech of various age groups: Data from Hungarian“. In: *Govor* 28.2, pp. 95–115 (cit. on pp. 73, 112).

Bona, Judit (2014). „Temporal characteristics of speech: The effect of age and speech style“. In: *The Journal of the Acoustical Society of America* 136.2, EL116–EL121 (cit. on pp. 73, 112).

Borensztajn, Gideon, Willem Zuidema, and Rens Bod (2009). „Childrens grammars grow more abstract with ageEvidence from an automatic procedure for identifying the productive units of language“. In: *Topics in Cognitive Science* 1.1, pp. 175–188 (cit. on p. 45).

Bosker, Hans Rutger, Anne-France Pinget, Hugo Quené, Ted Sanders, and Nivja H De Jong (2013). „What makes speech sound fluent? The contributions of pauses, speed and repairs“. In: *Language Testing* 30.2, pp. 159–175 (cit. on pp. 3, 40, 113).

Bosker, Hans Rutger, Hugo Quené, Ted Sanders, and Nivja H De Jong (2014a). „Native ums elicit prediction of low-frequency referents, but non-native ums do not“. In: *Journal of memory and language* 75, pp. 104–116 (cit. on p. 3).

– (2014b). „The perception of fluency in native and nonnative speech“. In: *Language Learning* 64.3, pp. 579–614 (cit. on pp. 40, 121).

Bosker, Hans Rutger, Eva Reinisch, and Matthias J Sjerps (2017). „Cognitive load makes speech sound fast, but does not modulate acoustic context effects“. In: *Journal of Memory and Language* 94, pp. 166–176 (cit. on p. 57).

Brennan, Susan E and Maurice Williams (1995). „The feeling of another s knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers“. In: *Journal of memory and language* 34.3, pp. 383–398 (cit. on p. 122).

Brevi, Matteo and Michael Ramscar (Oct. 2022). „unknown“. M.A. Thesis. University of Tuebingen (cit. on p. 129).

Buhusi, Catalin V and Warren H Meck (2005). „What makes us tick? Functional and neural mechanisms of interval timing“. In: *Nature reviews neuroscience* 6.10, pp. 755–765 (cit. on p. 76).

Buonomano, Dean V, Jennifer Bramen, and Mahsa Khodadadifar (2009). „Influence of the interstimulus interval on temporal processing and learning: testing the state-dependent network model“. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 364.1525, pp. 1865–1873 (cit. on p. 77).

Bürki, Audrey (2018). „Variation in the speech signal as a window into the cognitive architecture of language production“. In: *Psychonomic bulletin & review* 25.6, pp. 1973–2004 (cit. on pp. 3, 114).

Buz, Esteban, Michael K Tanenhaus, and T Florian Jaeger (2016). „Dynamically adapted context-specific hyper-articulation: Feedback from interlocutors affects speakers subsequent pronunciations“. In: *Journal of memory and language* 89, pp. 68–86 (cit. on p. 17).

Bybee, Joan (1985). „Diagrammatic iconicity in stem-inflection relations“. In: *Iconicity in syntax*, pp. 11–48 (cit. on p. 134).

Byrd, Dani (1994). „Relations of sex and dialect to reduction“. In: *Speech communication* 15.1-2, pp. 39–54 (cit. on p. 3).

Campbell, Karen L, Dávid Samu, Simon W Davis, et al. (2016). „Robust resilience of the frontotemporal syntax system to aging“. In: *Journal of Neuroscience* 36.19, pp. 5214–5227 (cit. on pp. 73, 89).

Campbell, Karen L and Lorraine K Tyler (2018). „Language-related domain-specific and domain-general systems in the human brain“. In: *Current opinion in behavioral sciences* 21, pp. 132–137 (cit. on p. 73).

Campione, Estelle and Jean Véronis (2002). „A large-scale multilingual study of silent pause duration“. In: *Speech prosody 2002, international conference* (cit. on p. 76).

Cancho, Ramon Ferrer-i (2017). „The placement of the head that maximizes predictability. An information theoretic approach“. In: *arXiv preprint arXiv:1705.09932* (cit. on p. 14).

Cancho, Ramon Ferrer-i, Antoni Hernández-Fernández, David Lusseau, et al. (2013). „Compression as a universal principle of animal behavior“. In: *Cognitive Science* 37.8, pp. 1565–1578 (cit. on p. 14).

Cancho, Ramon Ferrer I and Ricard V Solé (2003). „Least effort and the origins of scaling in human language“. In: *Proceedings of the National Academy of Sciences* 100.3, pp. 788–791 (cit. on p. 14).

Cao, Rosa (2020). „New labels for old ideas: Predictive processing and the interpretation of neural signals“. In: *Review of Philosophy and Psychology* 11.3, pp. 517–546 (cit. on p. 17).

Cao, Rosa and Daniel Yamins (2021). „Explanatory models in neuroscience: Part 1–taking mechanistic abstraction seriously“. In: *arXiv preprint arXiv:2104.01490* (cit. on p. 17).

Cervantes Constantino, Francisco and Jonathan Z Simon (2017). „Dynamic cortical representations of perceptual filling-in for missing acoustic rhythm“. In: *Scientific reports* 7.1, pp. 1–10 (cit. on p. 122).

Clark, Herbert H and Susan E Brennan (1991). „Grounding in communication.“ In: (cit. on p. 4).

Clayards, Meghan, Michael K Tanenhaus, Richard N Aslin, and Robert A Jacobs (2008). „Perception of speech reflects optimal use of probabilistic speech cues“. In: *Cognition* 108.3, pp. 804–809 (cit. on pp. 6, 70).

Clink, Dena J, Abdul Hamid Ahmad, and Holger Klinck (2020). „Brevity is not a universal in animal communication: evidence for compression depends on the unit of analysis in small ape vocalizations“. In: *Royal Society open science* 7.4, p. 200151 (cit. on p. 14).

Clopper, Cynthia G and Janet B Pierrehumbert (2008). „Effects of semantic predictability and regional dialect on vowel space reduction“. In: *The Journal of the Acoustical Society of America* 124.3, pp. 1682–1688 (cit. on p. 16).

Collins, Michael (2002). „Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms“. In: *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, pp. 1–8 (cit. on p. 45).

Cooke, Martin, Catherine Mayo, and Julián Villegas (2014). „The contribution of durational and spectral changes to the Lombard speech intelligibility benefit“. In: *The Journal of the Acoustical Society of America* 135.2, pp. 874–883 (cit. on pp. 40, 121).

Corley, Martin, Lucy J MacGregor, and David I Donaldson (2007). „Its the way that you, er, say it: Hesitations in speech affect language comprehension“. In: *Cognition* 105.3, pp. 658–668 (cit. on p. 113).

Coull, Jennifer T, Ruey-Kuang Cheng, and Warren H Meck (2011). „Neuroanatomical and neurochemical substrates of timing“. In: *Neuropsychopharmacology* 36.1, pp. 3–25 (cit. on p. 76).

Crystal, Thomas H and Arthur S House (1990). „Articulation rate and the duration of syllables and stress groups in connected speech“. In: *The Journal of the Acoustical Society of America* 88.1, pp. 101–112 (cit. on p. 124).

Cutler, Anne and Charles Clifton (1999). „Comprehending spoken language: A blueprint of the listener“. In: *The neurocognition of language*, pp. 123–166 (cit. on p. 70).

Dalal, Navneet and Bill Triggs (2005). „Histograms of oriented gradients for human detection“. In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. Vol. 1. IEEE, pp. 886–893 (cit. on p. 149).

Dautriche, Isabelle, Kyle Mahowald, Edward Gibson, and Steven T Piantadosi (2017). „Wordform similarity increases with semantic similarity: An analysis of 100 languages“. In: *Cognitive science* 41.8, pp. 2149–2169 (cit. on p. 14).

Davies, Mark (2010). „The Corpus of Contemporary American English as the first reliable monitor corpus of English“. In: *Literary and linguistic computing* 25.4, pp. 447–464 (cit. on pp. 32, 53, 119).

De Boer, Bart (2017). „Evolution of speech and evolution of language“. In: *Psychonomic bulletin & review* 24.1, pp. 158–162 (cit. on p. 143).

DeDeo, Simon (2018). „Information theory for intelligent people“. In: *Santa Fe* (cit. on pp. 20, 110).

DeDeo, Simon, Robert XD Hawkins, Sara Klingenstein, and Tim Hitchcock (2013). „Bootstrap methods for the empirical study of decision-making and information flows in social systems“. In: *Entropy* 15.6, pp. 2246–2276 (cit. on p. 20).

Dell, Gary S (1984). „Representation of serial order in speech: evidence from the repeated phoneme effect in speech errors.“ In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 10.2, p. 222 (cit. on pp. 1, 15).

Demberg, Vera and Frank Keller (2008). „Data from eye-tracking corpora as evidence for theories of syntactic processing complexity“. In: *Cognition* 109.2, pp. 193–210 (cit. on p. 123).

Demol, Mike, Werner Verhelst, and Piet Verhoeve (2006). „A study of speech pauses for multilingual time-scaling applications“. In: *Multilingual Speech and Language Processing* (cit. on pp. 76, 112).

DeRose, Steven J (1988). „Grammatical category disambiguation by statistical optimization“. In: *Computational linguistics* 14.1, pp. 31–39 (cit. on p. 45).

Diachek, Evgeniia and Sarah Brown-Schmidt (2022). „The effect of disfluency on memory for what was said.“ In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* (cit. on pp. 40, 113, 121).

Diedrichsen, Jörn, Yasmin Hashambhoy, Tushar Rane, and Reza Shadmehr (2005). „Neural correlates of reach errors“. In: *Journal of Neuroscience* 25.43, pp. 9919–9931 (cit. on p. 76).

Dienes, Zoltan (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. Springer (cit. on p. 114).

Dilley, Laura C and Mark A Pitt (2010). „Altering context speech rate can cause words to appear or disappear“. In: *Psychological Science* 21.11, pp. 1664–1670 (cit. on pp. 6, 72, 77, 105).

Dilts, Philip C (2013). „Modelling phonetic reduction in a corpus of spoken English using Random Forests and Mixed-Effects Regression“ (cit. on pp. 3, 42).

Doelling, Keith B, Luc H Arnal, Oded Ghitza, and David Poeppel (2014). „Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing“. In: *Neuroimage* 85, pp. 761–768 (cit. on pp. 88, 105).

Dominguez, Melissa and Robert A Jacobs (2003). „Developmental constraints aid the acquisition of binocular disparity sensitivities“. In: *Neural Computation* 15.1, pp. 161–182 (cit. on p. 104).

Duffy, Susan A and David B Pisoni (1992). „Comprehension of synthetic speech produced by rule: A review and theoretical interpretation“. In: *Language and Speech* 35.4, pp. 351–389 (cit. on p. 2).

Dye, Melody, Petar Milin, Richard Futrell, and Michael Ramscar (2018). „Alternative solutions to a language design problem: The role of adjectives and gender marking in efficient communication“. In: *Topics in cognitive science* 10.1, pp. 209–224 (cit. on pp. 19, 41, 130).

Edeline, Jean-Marc and Norman M Weinberger (1993). „Receptive field plasticity in the auditory cortex during frequency discrimination training: selective retuning independent of task difficulty.“ In: *Behavioral neuroscience* 107.1, p. 82 (cit. on p. 75).

Ellis, Nick C and Nuria Sagarra (2010). „The bounds of adult language acquisition: Blocking and learned attention“. In: *Studies in Second Language Acquisition* 32.4, pp. 553–580 (cit. on p. 35).

Elman, Jeffrey L (1993). „Learning and development in neural networks: The importance of starting small“. In: *Cognition* 48.1, pp. 71–99 (cit. on p. 104).

Ernestus, Mirjam, Harald Baayen, and Rob Schreuder (2002). „The recognition of reduced word forms“. In: *Brain and language* 81.1-3, pp. 162–173 (cit. on pp. 5, 6, 37, 70–72).

Estoup, Jean-Baptiste (1916). *Gammes sténographiques: méthode et exercices pour l'acquisition de la vitesse*. Institut sténographique (cit. on pp. 71, 146).

Evrard, Muriel (2002). „Ageing and lexical access to common and proper names in picture naming“. In: *Brain and Language* 81.1-3, pp. 174–179 (cit. on p. 115).

Fedzechkina, Maryia, T Florian Jaeger, and Elissa L Newport (2012). „Language learners restructure their input to facilitate efficient communication“. In: *Proceedings of the National Academy of Sciences* 109.44, pp. 17897–17902 (cit. on p. 14).

Fernald, Anne (1989). „Intonation and communicative intent in mothers' speech to infants: Is the melody the message?" In: *Child development*, pp. 1497–1510 (cit. on pp. 74, 104).

Finn, Amy S and Carla L Hudson Kam (2015). „Why segmentation matters: Experience-driven segmentation errors impair morpheme learning." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 41.5, p. 1560 (cit. on pp. 75, 105).

Finn, Amy S, Taraz Lee, Allison Kraus, and Carla L Hudson Kam (2014). „When it hurts (and helps) to try: The role of effort in language learning". In: *PloS one* 9.7, e101806 (cit. on p. 104).

Flack, Jessica C, Doug Erwin, Tanya Elliot, and David C Krakauer (2013). „Timescales, symmetry, and uncertainty reduction in the origins of hierarchy in biological systems". In: *Evolution cooperation and complexity*, pp. 45–74 (cit. on p. 159).

Fletcher, Annalise R, Megan J McAuliffe, Kaitlin L Lansford, and Julie M Liss (2015). „The relationship between speech segment duration and vowel centralization in a group of older speakers". In: *The Journal of the Acoustical Society of America* 138.4, pp. 2132–2139 (cit. on pp. 74, 112).

Foerde, Karin and Daphna Shohamy (2011). „Feedback timing modulates brain systems for learning in humans". In: *Journal of Neuroscience* 31.37, pp. 13157–13167 (cit. on p. 76).

Fors, Kristina Lundholm (2015). „Production and perception of pauses in speech". PhD thesis. Department of Philosophy, Linguistics, and Theory of Science, University of of Gothenburg (cit. on p. 76).

Fougeron, Cécile and Patricia A Keating (1997). „Articulatory strengthening at edges of prosodic domains". In: *The journal of the acoustical society of America* 101.6, pp. 3728–3740 (cit. on p. 57).

Fowler, Carol A (1992). „Vowel duration and closure duration in voiced and unvoiced stops: There are no contrast effects here". In: *Journal of Phonetics* 20.1, pp. 143–165 (cit. on p. 114).

Frantzi, Ilia and Michael Ramscar (Aug. 2022). „The Structure of Greek Gender Classes and how They Smooth Signalling in Noun Phrases". B.Sc. Thesis. University of Tuebingen (cit. on p. 129).

Fraundorf, Scott H and Duane G Watson (2011). „The disfluent discourse: Effects of filled pauses on recall". In: *Journal of memory and language* 65.2, pp. 161–175 (cit. on pp. 3, 40, 113).

Friberg, Anders and Johan Sundberg (1995). „Time discrimination in a monotonic, isochronous sequence". In: *The Journal of the Acoustical Society of America* 98.5, pp. 2524–2531 (cit. on pp. 74, 77).

Fuster, Joaquin M (2002). „Frontal lobe and cognitive development". In: *Journal of neurocytology* 31.3, pp. 373–385 (cit. on p. 104).

Futrell, Richard, Kyle Mahowald, and Edward Gibson (2015). „Large-scale evidence of dependency length minimization in 37 languages". In: *Proceedings of the National Academy of Sciences* 112.33, pp. 10336–10341 (cit. on p. 14).

Gahl, Susanne and R Harald Baayen (2019). „Twenty-eight years of vowels: Tracking phonetic variation through young to middle age adulthood". In: *Journal of Phonetics* 74, pp. 42–54 (cit. on pp. 74, 112).

Gahl, Susanne, Yao Yao, and Keith Johnson (2012). „Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech“. In: *Journal of memory and language* 66.4, pp. 789–806 (cit. on p. 1).

Genzel, Dmitriy and Eugene Charniak (2002). „Entropy rate constancy in text“. In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 199–206 (cit. on pp. 50, 82).

– (2003). „Variation of entropy and parse trees of sentences as a function of the sentence number“. In: *Proceedings of the 2003 conference on empirical methods in natural language processing*, pp. 65–72 (cit. on pp. 82, 158).

Gerstenberg, Annette, Susanne Fuchs, Julie Marie Kairet, Claudia Frankenberg, and Johannes Schröder (2018). „A cross-linguistic, longitudinal case study of pauses and interpausal units in spontaneous speech corpora of older speakers of German and French“. In: *group* 70, t3 (cit. on pp. 112, 114).

Ghitza, Oded (2011). „Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm“. In: *Frontiers in psychology* 2, p. 130 (cit. on p. 77).

Gibson, Edward, Richard Futrell, Steven P Piantadosi, et al. (2019). „How efficiency shapes human language“. In: *Trends in cognitive sciences* 23.5, pp. 389–407 (cit. on pp. 14, 160).

Gil, David (2008). „How much grammar does it take to sail a boat?(Or, what can material artefacts tell us about the evolution of language?)“ In: *The Evolution of Language*. World Scientific, pp. 123–130 (cit. on p. 45).

Gildea, Daniel and T Florian Jaeger (2015). „Human languages order information efficiently“. In: *arXiv preprint arXiv:1510.02823* (cit. on p. 14).

Giraud, Anne-Lise and David Poeppel (2012). „Cortical oscillations and speech processing: emerging computational principles and operations“. In: *Nature neuroscience* 15.4, pp. 511–517 (cit. on p. 107).

Gleitman, Lila R (2002). „Verbs of a feather flock together II The childs discovery of words and their meanings“. In: *The Legacy of Zellig Harris: Language and information into the 21st century. Volume 1: Philosophy of science, syntax and semantics* 228, p. 209 (cit. on pp. 5, 71).

Goh, K-I and A-L Barabási (2008). „Burstiness and memory in complex systems“. In: *EPL (Europhysics Letters)* 81.4, p. 48002 (cit. on p. 81).

Goldinger, Stephen D (1998). „Echoes of echoes? An episodic theory of lexical access.“ In: *Psychological review* 105.2, p. 251 (cit. on p. 16).

Goldman-Eisler, Frieda (1961). „The significance of changes in the rate of articulation“. In: *Language and Speech* 4.3, pp. 171–174 (cit. on p. 70).

Grabe, Esther and Ee Ling Low (2008). „Durational variability in speech and the rhythm class hypothesis“. In: *Laboratory phonology 7*. De Gruyter Mouton, pp. 515–546 (cit. on pp. 92, 93).

Grainger, Jonathan, Stéphane Dufau, Marie Montant, Johannes C Ziegler, and Joël Fagot (2012). „Orthographic processing in baboons (Papio papio)“. In: *Science* 336.6078, pp. 245–248 (cit. on pp. 142, 144, 148).

Gregory, Michelle L, William D Raymond, Alan Bell, Eric Fosler-Lussier, and Daniel Jurafsky (1999). „The effects of collocational strength and contextual predictability in lexical production“. In: *Chicago Linguistic Society*. Vol. 35. Citeseer, pp. 151–166 (cit. on p. 1).

Grice, H Paul (1969). „Utterer's meaning and intentions“. In: *The philosophical review* 78.2, pp. 147–177 (cit. on p. 4).

Grondin, Simon, Nicolas Bisson, and Caroline Gagnon (2011). „Sensitivity to time interval changes in speech and tone conditions“. In: *Attention, Perception, & Psychophysics* 73.3, pp. 720–728 (cit. on p. 76).

Groppe, David M, Marvin Choi, Tiffany Huang, et al. (2010). „The phonemic restoration effect reveals pre-N400 effect of supportive sentence context in speech perception“. In: *Brain research* 1361, pp. 54–66 (cit. on p. 122).

Grosjean, François (1980). „Spoken word recognition processes and the gating paradigm“. In: *Perception & psychophysics* 28.4, pp. 267–283 (cit. on p. 62).

Grosjean, Francois, Lysiane Grosjean, and Harlan Lane (1979). „The patterns of silence: Performance structures in sentence production“. In: *Cognitive psychology* 11.1, pp. 58–81 (cit. on p. 78).

Grosjean, François and Janna Itzler (1984). „Can semantic constraint reduce the role of word frequency during spoken-word recognition?“ In: *Bulletin of the Psychonomic Society* 22.3, pp. 180–182 (cit. on p. 62).

Hailman, Jack P, Millicent S Ficken, and Robert W Ficken (1985). „The chick-a-deecalls of Parus atricapillus: a recombinant system of animal communication compared with written English“. In: *Semiotica* 56.3-4, pp. 191–224 (cit. on pp. 146, 151).

Hale, John (2001). „A probabilistic Earley parser as a psycholinguistic model“. In: *Second meeting of the north american chapter of the association for computational linguistics* (cit. on pp. 20, 123).

– (2003). „The information conveyed by words in sentences“. In: *Journal of Psycholinguistic Research* 32.2, pp. 101–123 (cit. on p. 14).

– (2006). „Uncertainty about the rest of the sentence“. In: *Cognitive science* 30.4, pp. 643–672 (cit. on p. 14).

Hall, Kathleen Currie, Elizabeth Hume, T Florian Jaeger, and Andrew Wedel (2018). „The role of predictability in shaping phonological patterns“. In: *Linguistics vanguard* 4.s2 (cit. on pp. 3, 70, 160).

Harris, Zellig S (1954). „Distributional structure“. In: *Word* 10.2-3, pp. 146–162 (cit. on pp. 5, 71).

Hartley, Ralph VL (1928). „Transmission of information 1“. In: *Bell System technical journal* 7.3, pp. 535–563 (cit. on pp. 21, 33).

Hartshorne, Joshua K, Joshua B Tenenbaum, and Steven Pinker (2018). „A critical period for second language acquisition: Evidence from 2/3 million English speakers“. In: *Cognition* 177, pp. 263–277 (cit. on p. 104).

Hastie, Trevor and Robert Tibshirani (1990). „Exploring the nature of covariate effects in the proportional hazards model“. In: *Biometrics*, pp. 1005–1016 (cit. on p. 44).

Hazan, Valerie and Michèle Pettinato (2014). „The emergence of rhythmic strategies for clarifying speech: variation of syllable rate and pausing in adults, children and teenagers". In: *Proceedings of the 10th international seminar on speech production*, pp. 178–181 (cit. on pp. 73, 112).

Heldner, Mattias and Jens Edlund (2010). „Pauses, gaps and overlaps in conversations". In: *Journal of Phonetics* 38.4, pp. 555–568 (cit. on p. 76).

Herman, James P, Mark R Harwood, and Josh Wallman (2009). „Saccade adaptation specific to visual context". In: *Journal of Neurophysiology* 101.4, pp. 1713–1721 (cit. on pp. 16, 18, 89).

Hofstadter, Douglas and David Moser (1989). „To ebb is human; to study error-making is cognitive science". In: (cit. on p. 4).

Holzgrefe-Lang, Julia, Caroline Wellmann, Barbara Höhle, and Isabell Wartenburger (2018). „Infants processing of prosodic cues: Electrophysiological evidence for boundary perception beyond pause detection". In: *Language and speech* 61.1, pp. 153–169 (cit. on p. 73).

Hoopen, Gert Ten, Takayuki Sasaki, Yoshitaka Nakajima, et al. (2006). „Time-shrinking and categorical temporal ratio perception: evidence for a 1: 1 temporal category". In: *Music Perception* 24.1, pp. 1–22 (cit. on p. 77).

Hunter, Eric J, Mara Kapsner-Smith, Patrick Pead, Megan Zito Engar, and Wesley R Brown (2012). „Age and Speech Production: A Longitudinal Study of 50 Years". In: *Journal of the American Geriatrics Society* 60.6, p. 1175 (cit. on p. 112).

Iliev, Rumen and Robert Axelrod (2016). „Does causality matter more now? Increase in the proportion of causal language in English texts". In: *Psychological science* 27.5, pp. 635–643 (cit. on p. 110).

Iversen, John R, Aniruddh D Patel, and Kengo Ohgushi (2008). „Perception of rhythmic grouping depends on auditory experience". In: *The Journal of the Acoustical Society of America* 124.4, pp. 2263–2271 (cit. on p. 96).

Jacewicz, Ewa, Robert Allen Fox, and Lai Wei (2010). „Between-speaker and within-speaker variation in speech tempo of American English". In: *The Journal of the Acoustical Society of America* 128.2, pp. 839–850 (cit. on pp. 73, 111).

Jaeger, T Florian and Harry Tily (2011). „On language utility: Processing complexity and communicative efficiency". In: *Wiley Interdisciplinary Reviews: Cognitive Science* 2.3, pp. 323–335 (cit. on p. 14).

Jaekel, Brittany N, Rochelle S Newman, and Matthew J Goupell (2018). „Age effects on perceptual restoration of degraded interrupted sentences". In: *The Journal of the Acoustical Society of America* 143.1, pp. 84–97 (cit. on p. 122).

Jones, Michael N, Melody Dye, and Brendan T Johns (2017). „Context as an organizing principle of the lexicon". In: *Psychology of learning and motivation*. Vol. 67. Elsevier, pp. 239–283 (cit. on p. 40).

Jurafsky, Daniel, Alan Bell, Michelle Gregory, and William D Raymond (2001). „Probabilistic relations between words: Evidence from reduction in lexical production". In: *Typological studies in language* 45, pp. 229–254 (cit. on p. 113).

Karsai, Márton, Hang-Hyun Jo, Kimmo Kaski, et al. (2018). *Bursty human dynamics*. Springer (cit. on p. 81).

Katz, Slava M (1996). „Distribution of content words and phrases in text and language modelling“. In: *Natural language engineering* 2.1, pp. 15–59 (cit. on p. 81).

Kemps, Rachèl JJK, Mirjam Ernestus, Robert Schreuder, and R Harald Baayen (2005). „Prosodic cues for morphological complexity: The case of Dutch plural nouns“. In: *Memory & cognition* 33.3, pp. 430–446 (cit. on pp. 39, 58, 113).

Kidd, Evan, Caroline Junge, Tara Spokes, Lauren Morrison, and Anne Cutler (2018). „Individual differences in infant speech segmentation: Achieving the lexical shift“. In: *Infancy* 23.6, pp. 770–794 (cit. on p. 32).

Kiesling, Scott, Laura Dilley, and William D Raymond (2006). „The variation in conversation (ViC) project: Creation of the Buckeye Corpus of Conversational Speech“. In: *Language Variation and Change*, pp. 55–97 (cit. on p. 83).

King, Adam and Andrew Wedel (2020). „Greater early disambiguating information for less-probable words: The lexicon is shaped by incremental processing“. In: *Open Mind* 4, pp. 1–12 (cit. on pp. 14, 113, 116, 191).

Kirjavainen, Minna, Ludivine Crible, and Kate Beeching (2022). „Can filled pauses be represented as linguistic items? Investigating the effect of exposure on the perception and production of um“. In: *Language and Speech* 65.2, pp. 263–289 (cit. on p. 122).

Kleinschmidt, Dave F and T Florian Jaeger (2015). „Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel.“ In: *Psychological review* 122.2, p. 148 (cit. on pp. 6, 70).

Klingenstein, Sara, Tim Hitchcock, and Simon DeDeo (2014). „The civilizing process in Londons Old Bailey“. In: *Proceedings of the National Academy of Sciences* 111.26, pp. 9419–9424 (cit. on pp. 24, 45, 110, 118, 192).

Krakauer, John W, Pietro Mazzoni, Ali Ghazizadeh, Roshni Ravindran, and Reza Shadmehr (2006). „Generalization of motor learning depends on the history of prior action“. In: *PLoS Biol* 4.10, e316 (cit. on pp. 18, 73).

Kuhl, Patricia K (1981). „Discrimination of speech by nonhuman animals“. In: *The Journal of the Acoustical Society of America* 70.2, pp. 340–349 (cit. on p. 143).

– (2004). „Early language acquisition: cracking the speech code“. In: *Nature reviews neuroscience* 5.11, pp. 831–843 (cit. on pp. 103, 105).

Kuhl, Patricia K, Erica Stevens, Akiko Hayashi, et al. (2006). „Infants show a facilitation effect for native language phonetic perception between 6 and 12 months“. In: *Developmental science* 9.2, F13–F21 (cit. on pp. 103, 105).

Kulick, Don (2017). „Human–animal communication“. In: *Annual Review of Anthropology* 46, pp. 357–378 (cit. on p. 143).

Lamekina, Yulia and Lars Meyer (2022). „Entrainment to speech prosody influences subsequent sentence comprehension“. In: *Language, Cognition and Neuroscience*, pp. 1–14 (cit. on pp. 72, 105).

Landauer, Thomas K and Susan T Dumais (1997). „A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge.“ In: *Psychological review* 104.2, p. 211 (cit. on pp. 5, 71).

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). „Deep learning“. In: *nature* 521.7553, pp. 436–444 (cit. on p. 146).

Levelt, Willem JM (1993). „Lexical access in speech production“. In: *Knowledge and language*. Springer, pp. 241–251 (cit. on p. 15).

Levin, Beth (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago press (cit. on p. 102).

– (1995). „English verb classes and alternations“. In: *A preliminary Investigation* 1 (cit. on pp. 36, 158).

Levy, Roger (2008). „Expectation-based syntactic comprehension“. In: *Cognition* 106.3, pp. 1126–1177 (cit. on pp. 14, 20, 110, 123, 130, 160).

Lewis, Penelope A and R Christopher Miall (2003). „Distinct systems for automatic and cognitively controlled time measurement: evidence from neuroimaging“. In: *Current opinion in neurobiology* 13.2, pp. 250–255 (cit. on p. 76).

Liberman, Alvin M, Franklin S Cooper, Donald P Shankweiler, and Michael Studdert-Kennedy (1967). „Perception of the speech code.“ In: *Psychological review* 74.6, p. 431 (cit. on p. 70).

Lieberman, Philip, Liane S Feldman, Stanley Aronson, and Elizabeth Engen (1989). „Sentence comprehension, syntax and vowel duration in aged people“. In: *Clinical Linguistics & Phonetics* 3.4, pp. 299–311 (cit. on p. 74).

Ling, Low Ee, Esther Grabe, and Francis Nolan (2000). „Q uantitative characterizations of speech rhythm: Syllable-timing in Singapore English“. In: *Language and speech* 43.4, pp. 377–401 (cit. on p. 92).

Linke, Maja, Franziska Bröker, Michael Ramscar, and Harald Baayen (2017). „Are baboons learning" orthographic" representations? Probably not“. In: *PloS one* 12.8, e0183876 (cit. on pp. 16, 144, 147).

Linke, Maja and Michael Ramscar (2020a). „How the Probabilistic Structure of Grammatical Context Shapes Speech“. In: *Entropy* 22.1, p. 90 (cit. on pp. 11, 146, 152).

– (2020b). „How the Probabilistic Structure of Grammatical Context Shapes Speech“. In: *Entropy* 22.1 (cit. on pp. 70, 72, 79, 85, 100).

Lund, Kevin and Curt Burgess (1996). „Producing high-dimensional semantic spaces from lexical co-occurrence“. In: *Behavior research methods, instruments, & computers* 28.2, pp. 203–208 (cit. on pp. 5, 71).

Luo, Huan and David Poeppel (2007). „Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex“. In: *Neuron* 54.6, pp. 1001–1010 (cit. on pp. 17, 80, 107).

MacGregor, Lucy J, Martin Corley, and David I Donaldson (2010). „Listening to the sound of silence: Disfluent silent pauses in speech have consequences for listeners“. In: *Neuropsychologia* 48.14, pp. 3982–3992 (cit. on p. 113).

Mahowald, Kyle, Isabelle Dautriche, Edward Gibson, and Steven T Piantadosi (2018). „Word forms are structured for efficient use“. In: *Cognitive science* 42.8, pp. 3116–3134 (cit. on p. 14).

Mahowald, Kyle, Evelina Fedorenko, Steven T Piantadosi, and Edward Gibson (2013). „Info/information theory: Speakers choose shorter words in predictive contexts“. In: *Cognition* 126.2, pp. 313–318 (cit. on pp. 14, 21, 134).

Mampe, Birgit, Angela D Friederici, Anne Christophe, and Kathleen Wermke (2009). „Newborns' cry melody is shaped by their native language“. In: *Current biology* 19.23, pp. 1994–1997 (cit. on pp. 18, 73).

Männel, Claudia and Angela D Friederici (2009). „Pauses and intonational phrasing: ERP studies in 5-month-old German infants and adults“. In: *Journal of Cognitive Neuroscience* 21.10, pp. 1988–2006 (cit. on pp. 73, 74, 104).

– (2011). „Intonational phrase structure processing at different stages of syntax acquisition: ERP studies in 2-, 3-, and 6-year-old children“. In: *Developmental science* 14.4, pp. 786–798 (cit. on pp. 74, 104).

– (2016). „Neural correlates of prosodic boundary perception in German preschoolers: If pause is present, pitch can go“. In: *Brain research* 1632, pp. 27–33 (cit. on p. 74).

Männel, Claudia, Christine S Schipke, and Angela D Friederici (2013). „The role of pause as a prosodic boundary marker: Language ERP studies in German 3-and 6-year-olds“. In: *Developmental cognitive neuroscience* 5, pp. 86–94 (cit. on pp. 74, 104).

Marien, Peter, Sebastiaan Engelborghs, Franco Fabbro, and Peter P De Deyn (2001). „The lateralized linguistic cerebellum: a review and a new hypothesis“. In: *Brain and language* 79.3, pp. 580–600 (cit. on p. 73).

Maslowski, Merel, Antje S Meyer, and Hans Rutger Bosker (2019). „Listeners normalize speech for contextual speech rate even without an explicit recognition task“. In: *The Journal of the Acoustical Society of America* 146.1, pp. 179–188 (cit. on p. 73).

Mattys, Sven L, Peter W Jusczyk, Paul A Luce, and James L Morgan (1999). „Phonotactic and prosodic effects on word segmentation in infants“. In: *Cognitive psychology* 38.4, pp. 465–494 (cit. on pp. 103, 105).

McCOWAN, BRENDA, Sean F Hanser, and Laurance R Doyle (1999). „Quantitative tools for comparing animal communication systems: information theory applied to bottlenose dolphin whistle repertoires“. In: *Animal behaviour* 57.2, pp. 409–419 (cit. on pp. 146, 151).

Mcdonald, Craig M, Erik K Henricson, R Ted Abresch, et al. (2013). „The 6-minute walk test and other endpoints in Duchenne muscular dystrophy: longitudinal natural history observations over 48 weeks from a multicenter study“. In: *Muscle & nerve* 48.3, pp. 343–356 (cit. on pp. 22, 25).

McDonald, Scott and Michael Ramscar (2001). „Testing the distributioanl hypothesis: The influence of context on judgements of semantic similarity“. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 23. 23 (cit. on pp. 5, 35, 71).

McQueen, James M and LC Dilley (2020). „Prosody and spoken-word recognition“. In: *The Oxford handbook of language prosody*. Oxford University Press, pp. 509–521 (cit. on p. 100).

Mercado, Eduardo and Christina E Perazio (2022). „All units are equal in humpback whale songs, but some are more equal than others“. In: *Animal Cognition* 25.1, pp. 149–177 (cit. on pp. 14, 17).

Mercado III, Eduardo and Christina E Perazio (2021). „Similarities in composition and transformations of songs by humpback whales (Megaptera novaeangliae) over time and space.“ In: *Journal of Comparative Psychology* 135.1, p. 28 (cit. on p. 14).

Meyer, Lars (2018). „The neural oscillations of speech processing and language comprehension: state of the art and emerging mechanisms“. In: *European Journal of Neuroscience* 48.7, pp. 2609–2621 (cit. on p. 107).

Miller, Joanne L, François Grosjean, and Concetta Lomanto (1984). „Articulation rate and its variability in spontaneous speech: A reanalysis and some implications“. In: *Phonetica* 41.4, pp. 215–225 (cit. on pp. 70, 78).

Mitterer, Holger and James M McQueen (2009). „Processing reduced word-forms in speech perception using probabilistic knowledge about speech production.“ In: *Journal of Experimental Psychology: Human Perception and Performance* 35.1, p. 244 (cit. on p. 70).

Montemayor, Carlos and Marc Wittmann (2014). „The varieties of presence: Hierarchical levels of temporal integration“. In: *Timing & Time Perception* 2.3, pp. 325–338 (cit. on p. 121).

Moon-Hwan, Cho (2004). „Rhythm typology of Korean speech“. In: *Cognitive Processing* 5.4, pp. 249–253 (cit. on pp. 94, 95).

Morrill, Tuuli H, Laura C Dilley, J Devin McAuley, and Mark A Pitt (2014). „Distal rhythm influences whether or not listeners hear a word in continuous speech: Support for a perceptual grouping hypothesis“. In: *Cognition* 131.1, pp. 69–74 (cit. on pp. 6, 72, 75, 77, 105).

Mueller, Jutta L, Joerg Bahlmann, and Angela D Friederici (2008). „The role of pause cues in language learning: The emergence of event-related potentials related to sequence processing“. In: *Journal of Cognitive Neuroscience* 20.5, pp. 892–905 (cit. on p. 74).

Munson, Benjamin, Jan Edwards, Mary E Beckman, et al. (2011). „Phonological representations in language acquisition: Climbing the ladder of abstraction“. In: *Handbook of laboratory phonology*, pp. 288–309 (cit. on p. 74).

Munson, Benjamin, Jan Edwards, Sarah K Schellinger, Mary E Beckman, and Marie K Meyer (2010). „Deconstructing phonetic transcription: Covert contrast, perceptual bias, and an extraterrestrial view of Vox Humana“. In: *Clinical linguistics & phonetics* 24.4-5, pp. 245–260 (cit. on p. 32).

Nakajima, Yoshitaka, Gert Ten Hoopen, Gaston Hilkhuysen, and Takayuki Sasaki (1992). „Time-shrinking: A discontinuity in the perception of auditory temporal patterns“. In: *Perception & psychophysics* 51.5, pp. 504–507 (cit. on p. 77).

Nazzi, Thierry, Galina Iakimova, Josiane Bertoncini, Séverine Frédonie, and Carmela Alcantara (2006). „Early segmentation of fluent speech by infants acquiring French: Emerging evidence for crosslinguistic differences“. In: *Journal of Memory and Language* 54.3, pp. 283–299 (cit. on p. 32).

Neuberger, Tilda (2013). „Temporal patterns of childrens spontaneous speech“. In: *The Phonetician* 107.108, pp. 68–85 (cit. on pp. 73, 112).

Newman, Mark EJ (2005). „Power laws, Pareto distributions and Zipf's law“. In: *Contemporary physics* 46.5, pp. 323–351 (cit. on pp. 33, 34, 66).

Newport, Elissa L (1988). „Constraints on learning and their role in language acquisition: Studies of the acquisition of American Sign Language“. In: *Language sciences* 10.1, pp. 147–172 (cit. on p. 104).

Nixon, Jessie S (2020). „Of mice and men: Speech sound acquisition as discriminative learning from prediction error, not just statistical tracking“. In: *Cognition* 197, p. 104081 (cit. on p. 35).

Open Culture, LLC (2019). *How Magazine Pages Were Created Before Computers: A Veteran of the London Review of Books Demonstrates the Meticulous, Manual Process*. URL: \url{https://www.openculture.com/2019/10/how-magazine-pages-were-created-before-computers.html} (visited on Sept. 30, 2022) (cit. on pp. 27, 186).

Paris, Carol R, Margaret H Thomas, Richard D Gilson, and J Peter Kincaid (2000). „Linguistic cues and memory for synthetic and natural speech“. In: *Human Factors* 42.3, pp. 421–431 (cit. on p. 3).

Patel, Aniruddh D (2006). „Musical rhythm, linguistic rhythm, and human evolution“. In: *Music Perception* 24.1, pp. 99–104 (cit. on p. 75).

– (2021). „Vocal learning as a preadaptation for the evolution of human beat perception and synchronization“. In: *Philosophical Transactions of the Royal Society B* 376.1835, p. 20200326 (cit. on p. 75).

Peña, Marcela, Luca L Bonatti, Marina Nespor, and Jacques Mehler (2002). „Signal-driven computations in speech processing“. In: *Science* 298.5593, pp. 604–607 (cit. on p. 74).

Peters, Ole and Alexander Adamou (2022). „The ergodicity solution of the cooperation puzzle“. In: *Philosophical Transactions of the Royal Society A* 380.2227, p. 20200425 (cit. on pp. 79, 106).

Piantadosi, Steven T (2014). „Zipfs word frequency law in natural language: A critical review and future directions“. In: *Psychonomic bulletin & review* 21.5, pp. 1112–1130 (cit. on pp. 48, 49).

Piantadosi, Steven T, Harry Tily, and Edward Gibson (2011). „Word lengths are optimized for efficient communication“. In: *Proceedings of the National Academy of Sciences* 108.9, pp. 3526–3529 (cit. on pp. 14, 21, 58, 110, 160).

– (2012). „The communicative function of ambiguity in language“. In: *Cognition* 122.3, pp. 280–291 (cit. on pp. 14, 40).

Pitt, Mark A, Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond (2005). „The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability“. In: *Speech Communication* 45.1, pp. 89–95 (cit. on pp. 1, 41, 42, 46, 49, 64, 66, 83, 187, 197).

Poeppel, David and M Florencia Assaneo (2020). „Speech rhythms and their neural foundations“. In: *Nature Reviews Neuroscience* 21.6, pp. 322–334 (cit. on pp. 17, 80, 107).

Pollack, Irwin and JM Pickett (1963). „The intelligibility of excerpts from conversation“. In: *Language and Speech* 6.3, pp. 165–171 (cit. on p. 71).

Pöppel, Ernst (1997). „A hierarchical model of temporal perception“. In: *Trends in cognitive sciences* 1.2, pp. 56–61 (cit. on p. 121).

Pöppel, Ernst (2009). „Pre-semantically defined temporal windows for cognitive processing". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 364.1525, pp. 1887–1896 (cit. on p. 203).

Popper, Karl (2005). *The logic of scientific discovery*. Routledge (cit. on p. 38).

Port, Robert F and Adam P Leary (2005). „Against formal phonology". In: *Language* 81.4, pp. 927–964 (cit. on pp. 6, 31, 70, 72).

Poulisse, Charlotte, Linda Wheeldon, Rupali Limachya, Ali Mazaheri, and Katrien Segaert (2020). „The oscillatory mechanisms associated with syntactic binding in healthy ageing". In: *Neuropsychologia* 146, p. 107523 (cit. on p. 89).

Prince, Ellen F (1981). „Towards a taxonomy of given-new information". In: *Radical pragmatics* (cit. on p. 4).

Priva, Uriel Cohen (2008). „Using information content to predict phone deletion". In: *Proceedings of the 27th west coast conference on formal linguistics*. Cascadilla Proceedings Project Somerville, MA, pp. 90–98 (cit. on p. 110).

Priva, Uriel Cohen and T Florian Jaeger (2018). „The interdependence of frequency, predictability, and informativity in the segmental domain". In: *Linguistics Vanguard* 4.s2 (cit. on pp. 3, 70).

Quené, Hugo (2005). „Modeling of between-speaker and within-speaker variation in spontaneous speech tempo". In: *Ninth European Conference on Speech Communication and Technology* (cit. on pp. 73, 112, 124).

– (2007). „On the just noticeable difference for tempo in speech". In: *Journal of Phonetics* 35.3, pp. 353–362 (cit. on p. 74).

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria (cit. on p. 202).

Ramscar, Michael (2002). „The role of meaning in inflection: Why the past tense does not require a rule". In: *Cognitive Psychology* 45.1, pp. 45–94 (cit. on p. 134).

– (2019). „Source codes in human communication". In: *arXiv preprint arXiv:1904.03991* (cit. on pp. 3, 5, 7, 13, 19, 22, 24, 32–36, 43, 45, 47–49, 66, 70–72, 79, 100, 106, 146, 152, 158).

– (2020). „The empirical structure of word frequency distributions". In: *arXiv preprint arXiv:2001.05292* (cit. on p. 106).

– (2021a). „A discriminative account of the learning, representation and processing of inflection systems". In: *Language, Cognition and Neuroscience*, pp. 1–25 (cit. on pp. 3, 7, 19, 99, 100).

– (2021b). „How children learn to communicate discriminatively". In: *Journal of Child Language* 48.5, pp. 984–1022 (cit. on pp. 19, 72, 82).

Ramscar, Michael, Melody Dye, James Blevins, and Harald Baayen (2018). „Morphological development". In: *Handbook of Communications Disorders: Theoretical, Empirical, and Applied Linguistic Perspectives*, pp. 181–202 (cit. on pp. 6, 19, 72, 113).

Ramscar, Michael, Melody Dye, Jessica W Gustafson, and Joseph Klein (2013a). „Dual routes to cognitive flexibility: Learning and response-conflict resolution in the Dimensional Change Card Sort task". In: *Child development* 84.4, pp. 1308–1323 (cit. on p. 104).

Ramscar, Michael, Melody Dye, and Joseph Klein (2013b). „Children value informativity over logic in word learning“. In: *Psychological science* 24.6, pp. 1017–1023 (cit. on pp. 8, 104).

Ramscar, Michael, Melody Dye, and Stewart M McCauley (2013c). „Error and expectation in language learning: The curious absence of" mouses" in adult speech“. In: *Language*, pp. 760–793 (cit. on pp. 8, 19, 104, 146).

Ramscar, Michael and Nicole Gitcho (2007). „Developmental change and the nature of learning in childhood“. In: *Trends in cognitive sciences* 11.7, pp. 274–279 (cit. on pp. 19, 104, 147).

Ramscar, Michael, Peter Hendrix, Bradley Love, and R Harald Baayen (2013d). „Learning is not decline: The mental lexicon as a window into cognition across the lifespan“. In: *The Mental Lexicon* 8.3, pp. 450–481 (cit. on pp. 74, 84).

Ramscar, Michael, Peter Hendrix, Cyrus Shaoul, Petar Milin, and Harald Baayen (2014). „The myth of cognitive decline: Non-linear dynamics of lifelong learning“. In: *Topics in cognitive science* 6.1, pp. 5–42 (cit. on pp. 8, 32, 71, 81, 82, 89, 99, 111, 115, 117, 192).

Ramscar, Michael and Robert F Port (2016). „How spoken languages work in the absence of an inventory of discrete units“. In: *Language Sciences* 53, pp. 58–74 (cit. on pp. 5, 7, 13, 31, 32, 35).

Ramscar, Michael, Ching Chu Sun, Peter Hendrix, and Harald Baayen (2017). „The mismeasurement of mind: Life-span changes in paired-associate-learning scores reflect the cost of learning, not cognitive decline“. In: *Psychological science* 28.8, pp. 1171–1179 (cit. on pp. 8, 74, 111, 115, 117).

Ramscar, Michael and Daniel Yarlett (2007). „Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition“. In: *Cognitive science* 31.6, pp. 927–960 (cit. on pp. 8, 35).

Ramscar, Michael, Daniel Yarlett, Melody Dye, Katie Denny, and Kirsten Thorpe (2010). „The effects of feature-label-order and their implications for symbolic learning“. In: *Cognitive science* 34.6, pp. 909–957 (cit. on pp. 5, 35).

Ramscar, Michael et al. (2013e). „Suffixing, prefixing, and the functional order of regularities in meaningful strings“. In: *Psihologija* 46.4, pp. 377–396 (cit. on pp. 19, 100, 113, 116, 191).

Raymond, William D, Mark Pitt, Keith Johnson, et al. (2002). „An analysis of transcription consistency in spontaneous speech from the Buckeye corpus“. In: *Seventh International Conference on Spoken Language Processing* (cit. on p. 5).

Redford, Melissa A (2013). „A comparative analysis of pausing in child and adult storytelling“. In: *Applied Psycholinguistics* 34.3, p. 569 (cit. on pp. 73, 112).

Repp, Bruno H and Yi-Huang Su (2013). „Sensorimotor synchronization: a review of recent research (2006–2012)“. In: *Psychonomic bulletin & review* 20.3, pp. 403–452 (cit. on p. 77).

Rescorla, Robert A (1972). „A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement“. In: *Current research and theory*, pp. 64–99 (cit. on p. 148).

Rivera-Gaxiola, Maritza, Lindsay Klarman, Adrian Garcia-Sierra, and Patricia K Kuhl (2005). „Neural patterns to speech and vocabulary growth in American infants“. In: *NeuroReport* 16.5, pp. 495–498 (cit. on p. 103).

Rohde, Douglas LT and David C Plaut (1999). „Language acquisition in the absence of explicit negative evidence: How important is starting small?“ In: *Cognition* 72.1, pp. 67–109 (cit. on p. 104).

Sahlgren, Magnus (2008). „The distributional hypothesis“. In: *Italian Journal of Disability Studies* 20, pp. 33–53 (cit. on pp. 5, 71).

Saija, Jefta D, Elkan G Akyürek, Tjeerd C Andringa, and Deniz Başkent (2014). „Perceptual restoration of degraded speech is preserved with advancing age“. In: *Journal of the Association for Research in Otolaryngology* 15.1, pp. 139–148 (cit. on p. 122).

Salverda, Anne Pier, Delphine Dahan, and James M McQueen (2003). „The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension“. In: *Cognition* 90.1, pp. 51–89 (cit. on pp. 39, 57, 113).

Sampson, Geoffrey (2013). „A counterexample to homophony avoidance“. In: *Diachronica* 30.4, pp. 579–591 (cit. on pp. 15, 38).

– (2019). „An unaddressed phonological contradiction“. In: *International Journal of Chinese Linguistics* 6.2, pp. 221–237 (cit. on pp. 15, 25).

Samuel, Arthur G (2001). „Knowing a word affects the fundamental perception of the sounds within it“. In: *Psychological Science* 12.4, pp. 348–351 (cit. on p. 122).

– (2020). „Psycholinguists should resist the allure of linguistic units as perceptual units“. In: *Journal of Memory and Language* 111, p. 104070 (cit. on p. 70).

Sasaki, Takayuki, Yoshitaka Nakajima, Gert Ten Hoopen, et al. (2010). „Time stretching: Illusory lengthening of filled auditory durations“. In: *Attention, Perception, & Psychophysics* 72.5, pp. 1404–1421 (cit. on p. 77).

Scarf, Damian, Karoline Boy, Anelisie Uber Reinert, et al. (2016). „Orthographic processing in pigeons (Columba livia)“. In: *Proceedings of the National Academy of Sciences* 113.40, pp. 11272–11276 (cit. on p. 144).

Schachter, Stanley, Nicholas Christenfeld, Bernard Ravina, and Frances Bilous (1991). „Speech disfluency and the structure of knowledge.“ In: *Journal of personality and social psychology* 60.3, p. 362 (cit. on p. 3).

Schrödinger, Erwin (1935). „Discussion of probability relations between separated systems“. In: *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 31. 4. Cambridge University Press, pp. 555–563 (cit. on p. 22).

Scribner, Sylvia and Michael Cole (2013). *The psychology of literacy*. Harvard University Press (cit. on p. 32).

Seidl, Amanda and Alejandrina Cristià (2008). „Developmental changes in the weighting of prosodic cues“. In: *Developmental Science* 11.4, pp. 596–606 (cit. on p. 73).

Seifart, Frank, Jan Strunk, Swintha Danielsen, et al. (2018). „Nouns slow down speech across structurally and culturally diverse languages“. In: *Proceedings of the National Academy of Sciences* 115.22, pp. 5720–5725 (cit. on pp. 37, 115).

Seyfarth, Scott (2014). „Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation“. In: *Cognition* 133.1, pp. 140–155 (cit. on pp. 3, 70).

Seyfarth, Scott, Esteban Buz, and T Florian Jaeger (2016). „Dynamic hyperarticulation of coda voicing contrasts“. In: *The Journal of the Acoustical Society of America* 139.2, EL31–EL37 (cit. on pp. 3, 70).

Shannon, Claude E (1948). „A mathematical theory of communication“. In: *The Bell system technical journal* 27.3, pp. 379–423 (cit. on pp. 7, 14, 21, 33, 47, 70, 75, 108, 111).

Shatzman, Keren B and James M McQueen (2006). „Segment duration as a cue to word boundaries in spoken-word recognition“. In: *Perception & Psychophysics* 68.1, pp. 1–16 (cit. on p. 113).

Shmuelof, Lior and John W Krakauer (2014). „Recent insights into perceptual and motor skill learning“. In: *Frontiers in human neuroscience* 8, p. 683 (cit. on pp. 19, 89).

Shmuelof, Lior, John W Krakauer, and Pietro Mazzoni (2012). „How is a motor skill learned? Change and invariance at the levels of task success and trajectory control“. In: *Journal of neurophysiology* 108.2, pp. 578–594 (cit. on p. 19).

Sivonen, Päivi, Burkhard Maess, and Angela D Friederici (2006). „Semantic retrieval of spoken words with an obliterated initial phoneme in a sentence context“. In: *Neuroscience letters* 408.3, pp. 220–225 (cit. on p. 122).

Skoruppa, Katrin, Ferran Pons, Laura Bosch, et al. (2013). „The development of word stress processing in French and Spanish infants“. In: *Language Learning and Development* 9.1, pp. 88–104 (cit. on pp. 74, 104).

Smalle, Eleonore HM, Muriel Panouilleres, Arnaud Szmalec, and Riikka Möttönen (2017). „Language learning in the adult brain: Disrupting the dorsolateral prefrontal cortex facilitates word-form learning“. In: *Scientific reports* 7.1, pp. 1–9 (cit. on p. 105).

Sommers, Mitchell S and Stephanie M Danielson (1999). „Inhibitory processes and spoken word recognition in young and older adults: The interaction of lexical competition and semantic context.“ In: *Psychology and aging* 14.3, p. 458 (cit. on p. 114).

Soni, Sandeep, Lauren Klein, and Jacob Eisenstein (2021). „Abolitionist networks: Modeling language change in nineteenth-century activist newspapers“. In: *arXiv preprint arXiv:2103.07538* (cit. on p. 110).

Stanley, Jason and John W Krakauer (2013). „Motor skill depends on knowledge of facts“. In: *Frontiers in human neuroscience* 7, p. 503 (cit. on p. 18).

Stivers, Tanya, Nicholas J Enfield, Penelope Brown, et al. (2009). „Universals and cultural variation in turn-taking in conversation“. In: *Proceedings of the National Academy of Sciences* 106.26, pp. 10587–10592 (cit. on p. 72).

Stone, Gregory O (1986). „An analysis of the delta rule and the learning of statistical associations“. In: *Explorations in the microstructure of cognition* 1, pp. 444–459 (cit. on p. 148).

Sundara, Megha, Monika Molnar, and Sónia Frota (2015). „The perception of boundary tones in infancy“. In: *Proceedings of the 18th International Congress of Phonetic Sciences*. International Phonetic Association, pp. 1–4 (cit. on pp. 74, 104).

Sundara, Megha and Adrienne Scutellaro (2011). „Rhythmic distance between languages affects the development of speech perception in bilingual infants“. In: *Journal of Phonetics* 39.4, pp. 505–513 (cit. on pp. 74, 104).

Taleb, Nassim Nicholas (2007). „Black swans and the domains of statistics“. In: *The american statistician* 61.3, pp. 198–200 (cit. on p. 25).

Taleb, Nassim Nicholas, Yaneer Bar-Yam, and Pasquale Cirillo (2020). „On single point forecasts for fat-tailed variables“. In: *International Journal of Forecasting* (cit. on p. 25).

Tark, Eun-Sun (2012). „An experimental study of Korean rhythm structure on the basis of rhythm metrics“. In: *The Journal of the Acoustical Society of America* 132.3, pp. 2005–2005 (cit. on pp. 94, 95).

Teffer, Kate and Katerina Semendeferi (2012). „Human prefrontal cortex: evolution, development, and pathology“. In: *Progress in brain research* 195, pp. 191–218 (cit. on p. 104).

Teki, Sundeep, Manon Grube, Sukhbinder Kumar, and Timothy D Griffiths (2011). „Distinct neural substrates of duration-based and beat-based auditory timing“. In: *Journal of Neuroscience* 31.10, pp. 3805–3812 (cit. on p. 76).

Thompson-Schill, Sharon L, Michael Ramscar, and Evangelia G Chrysikou (2009). „Cognition without control: When a little frontal lobe goes a long way“. In: *Current directions in psychological science* 18.5, pp. 259–263 (cit. on p. 104).

Tily, Harry, Susanne Gahl, Inbal Arnon, et al. (2009). „Syntactic probabilities affect pronunciation variation in spontaneous speech“. In: *Language and Cognition* 1.2, pp. 147–165 (cit. on pp. 39, 40, 58, 70).

Tomaschek, Fabian, Denis Arnold, Franziska Bröker, and R Harald Baayen (2018). „Lexical frequency co-determines the speed-curvature relation in articulation“. In: *Journal of phonetics* 68, pp. 103–116 (cit. on p. 116).

Tremblay, Stéphanie, Guillaume Houle, and David J Ostry (2008). „Specificity of speech motor learning“. In: *Journal of Neuroscience* 28.10, pp. 2426–2434 (cit. on pp. 73, 116).

Tsao, Feng-Ming, Huei-Mei Liu, and Patricia K Kuhl (2006). „Perception of native and non-native affricate-fricative contrasts: Cross-language tests on adults and infants“. In: *The Journal of the Acoustical Society of America* 120.4, pp. 2285–2294 (cit. on p. 103).

Tucha, Oliver and Klaus W Lange (2004). „Handwriting and attention in children and adults with attention deficit hyperactivity disorder“. In: *Motor control* 8.4, pp. 461–471 (cit. on pp. 18, 89).

Tucker, Benjamin V, Catherine Ford, and Stephanie Hedges (2021). „Speech aging: Production and perception“. In: *Wiley Interdisciplinary Reviews: Cognitive Science* 12.5, e1557 (cit. on pp. 73, 113).

Tuller, Betty and JA Scott Kelso (2018). „Phase transitions in speech production and their perceptual consequences“. In: *Attention and performance XIII*. Psychology Press, pp. 429–452 (cit. on p. 112).

Van Son, RJJH, Louis CW Pols, et al. (2003a). „An acoustic model of communicative efficiency in consonants and vowels taking into account context distinctiveness“. In: *Proceedings of ICPhS2003, Barcelona, Spain* (cit. on pp. 38, 62).

– (2003b). „How efficient is speech“. In: *Proceedings of the institute of phonetic sciences*. Vol. 25, pp. 171–184 (cit. on pp. 57, 62).

Vogelsang, Marin, Lukas Vogelsang, Sidney Diamond, and Pawan Sinha (2022). „Prenatal auditory experience and its sequelae“. In: *Developmental Science*, e13278 (cit. on p. 104).

Walker, Michael B and Carmelina Trimboli (1982). „Smooth transitions in conversational interactions“. In: *The Journal of Social Psychology* 117.2, pp. 305–306 (cit. on p. 76).

Warren, Richard M (1970). „Perceptual restoration of missing speech sounds“. In: *Science* 167.3917, pp. 392–393 (cit. on pp. 70, 122).

Warren, Richard M and Gary L Sherman (1974). „Phonemic restorations based on subsequent context“. In: *Perception & Psychophysics* 16.1, pp. 150–156 (cit. on p. 122).

Watson, Peter J and Benjamin Munson (2007). „A comparison of vowel acoustics between older and younger adults“. In: *Proceedings of the 16th international congress of phonetic sciences*, pp. 561–564 (cit. on pp. 74, 112).

Wedel, Andrew (2012). „Lexical contrast maintenance and the organization of sublexical contrast systems“. In: *Language and Cognition* 4.4, pp. 319–355 (cit. on pp. 15, 38, 110).

Wedel, Andrew, Scott Jackson, and Abby Kaplan (2013a). „Functional load and the lexicon: Evidence that syntactic category and frequency relationships in minimal lemma pairs predict the loss of phoneme contrasts in language change“. In: *Language and speech* 56.3, pp. 395–417 (cit. on p. 40).

Wedel, Andrew, Abby Kaplan, and Scott Jackson (2013b). „High functional load inhibits phonological contrast loss: A corpus study“. In: *Cognition* 128.2, pp. 179–186 (cit. on pp. 3, 15, 58, 70, 110, 113).

Wedel, Andrew, Noah Nelson, and Rebecca Sharp (2018). „The phonetic specificity of contrastive hyperarticulation in natural speech“. In: *Journal of Memory and Language* 100, pp. 61–88 (cit. on pp. 3, 15, 40, 58, 70, 110, 113).

Wedel, Andrew, Adam Ussishkin, and Adam King (2019a). „Crosslinguistic evidence for a strong statistical universal: Phonological neutralization targets word-ends over beginnings“. In: *Language* 95.4, e428–e446 (cit. on p. 15).

Wedel, Andrew B, Kathleen Currie Hall, T Florian Jaeger, and Elizabeth Hume (2019b). „The Message Shapes Phonology“. In: (cit. on p. 15).

Weilhammer, Karl and Susen Rabold (2003). „Durational aspects in turn taking“. In: *Proceedings of the International Conference of Phonetic Sciences*, pp. 2145–2148 (cit. on p. 72).

Werker, Janet F and Richard C Tees (1984). „Cross-language speech perception: Evidence for perceptual reorganization during the first year of life“. In: *Infant behavior and development* 7.1, pp. 49–63 (cit. on pp. 103, 105).

Wermke, Kathleen, Michael P Robb, and Philip J Schluter (2021). „Melody complexity of infants cry and non-cry vocalisations increases across the first six months“. In: *Scientific reports* 11.1, pp. 1–11 (cit. on pp. 18, 73).

Wester, Mirjam, Oliver Watts, and Gustav Eje Henter (2016). „Evaluating comprehension of natural and synthetic conversational speech“. In: *Proc. speech prosody*. Vol. 8, pp. 736–740 (cit. on p. 2).

White, Peter A (2017). „The three-second subjective present: A critical review and a new proposal.“ In: *Psychological Bulletin* 143.7, p. 735 (cit. on pp. 121, 203).

Widrow, Bernard and Marcian E Hoff (1960). *Adaptive switching circuits*. Tech. rep. Stanford Electronics Labs (cit. on p. 148).

Wilson, Margaret and Thomas P Wilson (2005). „An oscillator model of the timing of turn-taking“. In: *Psychonomic bulletin & review* 12.6, pp. 957–968 (cit. on p. 72).

Winter, Bodo and Andrew Wedel (2016). „The Co-evolution of Speech and the Lexicon: The Interaction of Functional Pressures, Redundancy, and Category Variation“. In: *Topics in cognitive science* 8.2, pp. 503–513 (cit. on p. 15).

Wong, Aaron L, Jeff Goldsmith, Alexander D Forrence, Adrian M Haith, and John W Krakauer (2017). „Reaction times can reflect habits rather than computations“. In: *Elife* 6, e28075 (cit. on p. 73).

Wood, Simon N (2006). *Generalized additive models: an introduction with R*. chapman and hall/CRC (cit. on p. 44).

– (2017). *Generalized additive models: an introduction with R*. Chapman and Hall/CRC (cit. on pp. 59, 120, 127).

Woolf, Virginia and Nigel Nicolson (1975). *The Letters of Virginia Woolf*. Harvest Books (cit. on p. 107).

Xie, Xin and Emily B Myers (2017). „Learning a talker or learning an accent: Acoustic similarity constrains generalization of foreign accent adaptation to new talkers“. In: *Journal of Memory and Language* 97, pp. 30–46 (cit. on pp. 16, 105).

Xie, Xin, Rachel M Theodore, and Emily B Myers (2017). „More than a boundary shift: Perceptual adaptation to foreign-accented speech reshapes the internal structure of phonetic categories.“ In: *Journal of Experimental Psychology: Human Perception and Performance* 43.1, p. 206 (cit. on pp. 16, 75, 105).

Yan, Jun (2003). „Canadian Association of Neuroscience Review: development and plasticity of the auditory cortex“. In: *Canadian journal of neurological sciences* 30.3, pp. 189–200 (cit. on p. 73).

Yun, Weonhee, Kyuchul Yoon, Sunwoo Park, et al. (2015). „The Korean corpus of spontaneous speech“. In: *Phonetics and Speech Sciences* 7.2, pp. 103–109 (cit. on p. 83).

Ziegler, Johannes C, Thomas Hannagan, Stéphane Dufau, et al. (2013). „Transposed-letter effects reveal orthographic processing in baboons“. In: *Psychological science* 24.8, pp. 1609–1611 (cit. on pp. 142, 144, 148, 194).

Zimmer, Benjamin (2012). „New Microscopes, New Telescopes: A Conversation with Mark Liberman About the Uncertain Future of Linguistics“. In: *American speech* 87.1, pp. 107–108 (cit. on p. 145).

Zipf, George Kingsley (1949). *Human behavior and the principle of least effort: an introd. to human ecology* (cit. on pp. 14, 21, 71, 146).

List of Figures

List of Tables

Declaration

I declare that the thesis has been composed by myself and that the work has not be submitted for any other degree or professional qualification. I confirm that the work submitted is my own, except where work which has formed part of jointly-authored publications has been included. My contribution and those of the other authors to this work have been explicitly indicated below. I confirm that appropriate credit has been given within this thesis where reference has been made to the work of others.

Three of the studies reported in this thesis were a collaboration. For the published/-submitted projects (Chapter 3, 5 and 7), I played the principal role in conceptual genesis, data collection, analysis and visualization of the results and writing of the initial drafts. The final versions of the articles were written, and edited in collaboration with Michael Ramscar, I wrote approximately 80% of the text.

Tübingen, May 15, 2023

Maja Linke

Appendix

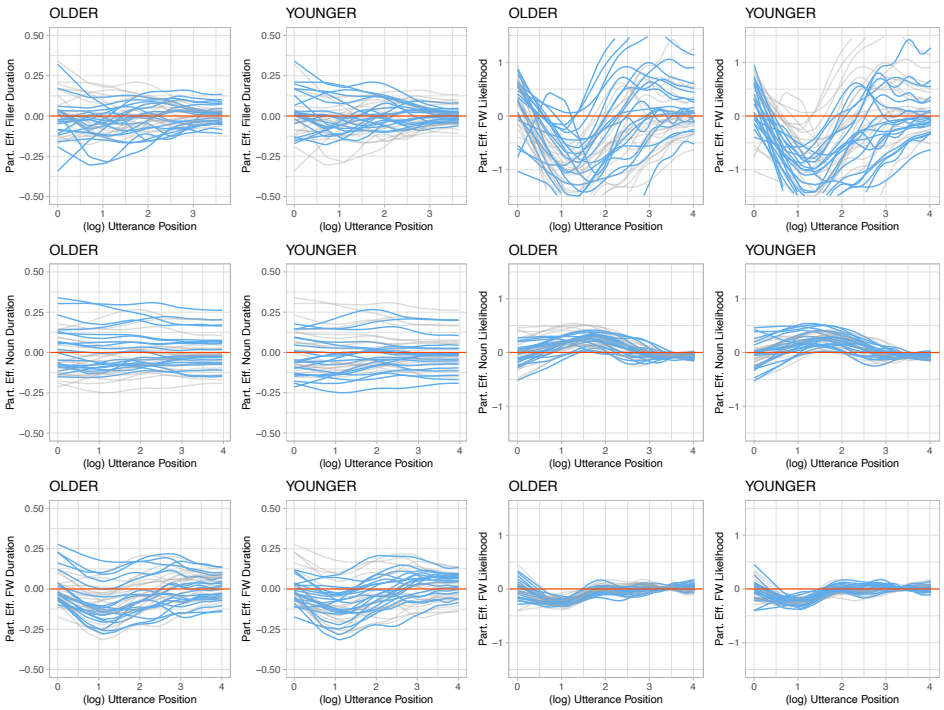

Fig. 8.1: Comparison plot, Ch. 6.5, factor smooths for individual speakers from different age cohorts:fillers (top), nouns (center), and function words (bottom). The target cohort is highlighted in blue. Left: Factor smooths for duration as a smooth function of utterance position. Right: Factor smooths fitted to the log odds of word/filler for utterance position in older speakers and younger speakers. Fillers and function word likelihood follow a similar pattern (u-shaped pattern in the left side of the plot), individual speakers smooths are much 'noisier' in fillers than in function words.

Chapter 5, Supplements

Distribution analysis: fitting, tests and descriptive statistics

This section contains the description of the fitting procedures for the data presented in the first result section of chapter 5 (the beginning of section 4 and subsection 4.1).

The analyses were conducted in the R environment version 4.2.0 (R Core Team, 2022). For distribution fitting we use the *fitdistr* function for maximum-likelihood fitting of univariate distributions from the *MASS* library (version 7.3-56). The truncated distributions were fitted with the *RelIns* library.

The density function of the exponential distribution with a rate λ is defined as

$$f_X(x) = \lambda \mathrm{e}^{-\lambda x} x \geq 00 x < 0 \tag{8.1}$$

From the fitted exponential distribution we extract the rate parameter λ and fit two models: an exponential model distribution with the random generator function *rexp* from the R *stats* library and a truncated exponential distribution for the selected range of values setting the endpoint parameter to 3000 ms with the *rtexp* function from the *RelIns* package (version 1.0.10). The truncated exponential distribution serves as the empirical model (because empirically, pauses and intervals longer than 3000 ms seem to produce qualitatively different responses). The exponential distribution fitted to the rate parameter serves as the theoretical model.

We tested the distance between the model exponential, the truncated model, and the empirical distribution with the two-sample KolmogorovSmirnov test. The KS statistic is a non-parametric test of equality between two continuous probability distributions. It quantifies the distance between the empirical sample and the cumulative function of the model distribution. The two-sample test is the most common method for comparing samples from continuous distributions. It is sensitive to differences in location and shape of the cumulative distribution functions.

Note that we only report fits to the exponential distribution in the main document. The fits of the three 'competing' distributions, Weibull, gamma, and log-normal, were also tested. The results are inconclusive. Independent of the procedures used to fit the parameters and the limits (i.e. $(0 - 2700)$, $(0 - 3000)$ or $(0 - \infty)$, the tests show small, significant differences between the model and the aggregate at our disposal.

The statistic seems to be relatively insensitive to the differences in the tail of the distribution.

Note that the distinctions between the distributions are motivated by theoretical definitions of limits (e.g. log-linear allows for negative reals while exponential, gamma, and Weibull do not), shape (exponential is the special case of Weibull and gamma when their shape parameter k equals 1) and scale parameters, which can be thought of as approximating how fast events occur in a fixed timeframe. Taken together, all of these distributions serve to distinguish between different ways in which observable classes of events (intervals or quantities) tend to behave in samples. The distinctions are somewhat arbitrary and often only reflect the properties of sampling; the differences between fits to Weibull, log-linear, gamma, and exponential distributions, seem to arbitrarily correlate to the limits set by the method of measurement, the variable nature of the data we are dealing with (annotated spontaneous speech), and the size of the sample.

The limits of the phenomenon examined in this work are set by the fact that pause durations always yield positive values, that longer intervals (of up to 3000 ms) appear to be bounded by the way people experience sensory input in time (Pöppel, 2009; White, 2017), and that pauses shorter than 50 ms are not reliably distinguished by speakers **and** annotation methods. Given all of the foregoing, and the relatively small distance between the truncated exponential model and the empirical distribution of pauses, we take that the evidence provided by our analyses and the test statistic (see Figure 8.2) are sufficient to support the hypothesis that the aggregate distribution of speech pause durations approximates the exponential distribution.

Tab. 8.1: Distribution of pause duration from samples of conversational English and Korean. Mean durations in random samples drawn from the empirical distribution of Korean pauses are less dispersed than the model mean distribution. As shown in Figure 1 of the results section of our article, samples of consecutive pauses from the empirical distribution are more dispersed than consecutive samples from the model distribution. This indicates that pauses from different ranges of duration are not uniformly distributed in time, and instead appear in local bursts.

	RANDOM					
ENGLISH	mean	median	SD	SE	IOD[1]	CV[2](%)
model	407.63	407.35	60.26	1.91	8.91	14.78
empirical	409.96	405.86	59.85	1.89	8.74	14.60
KOREAN						
model	468.13	463.06	67.96	2.15	9.87	14.51
empirical	464.14	461.25	51.89	1.64	**5.80**	11.18

	CONSECUTIVE					
ENGLISH	mean	median	SD	SE	IOD	CV(%)
model	412.40	409.51	54.77	1.73	7.27	13.28
empirical	412.12	390.90	126.83	4.01	**39.03**	30.78
KOREAN						
model	466.75	464.31	61.56	1.95	8.12	13.19
empirical	468.23	447.11	130.32	4.12	**36.27**	27.83

[1]Index of Dispersion, [2]Coefficient of Variance (%)

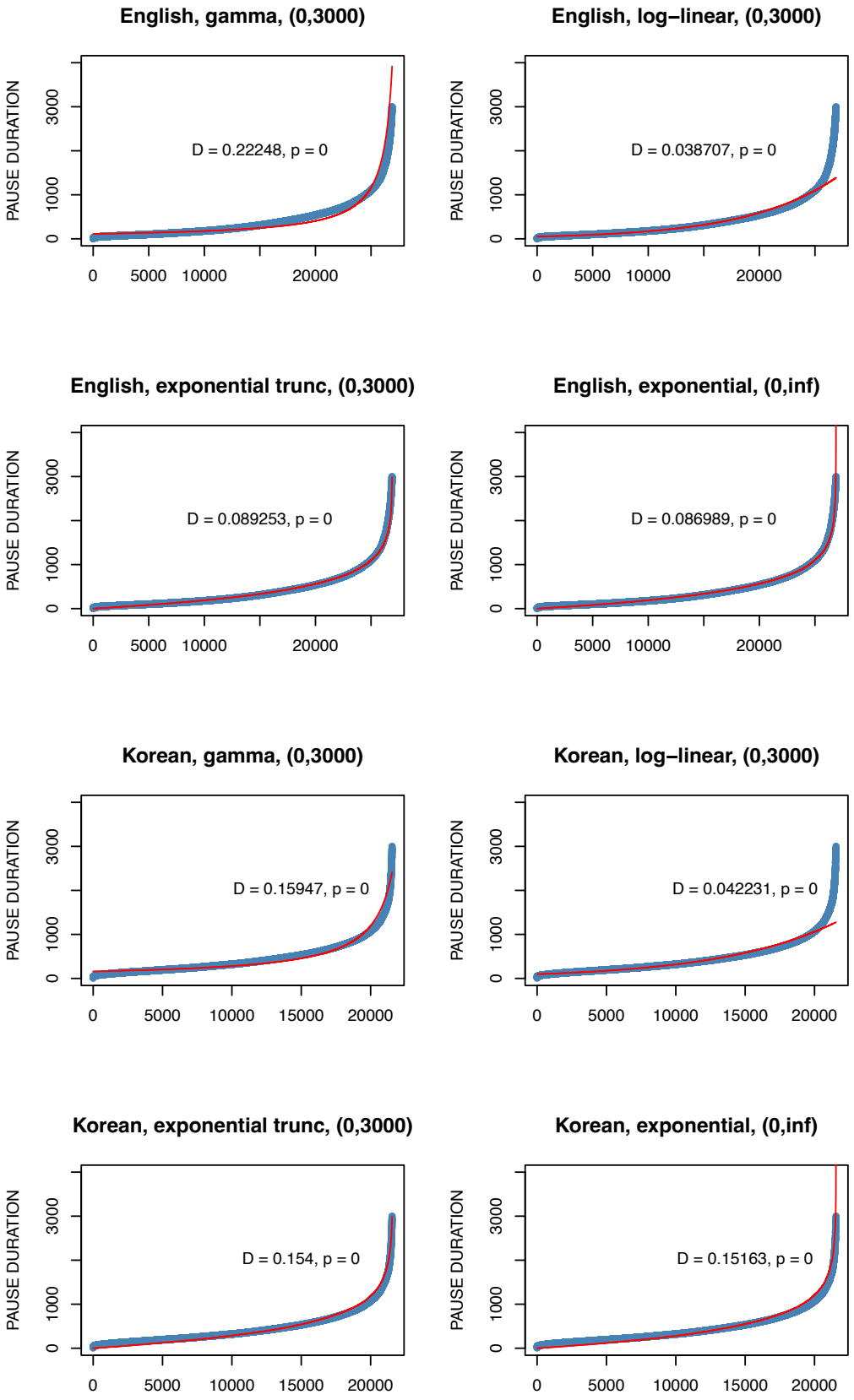

Fig. 8.2: The empirical distribution of pauses from the English and Korean samples plotted against the model distributions (red line). The Kolmogorov-Smirnov test statistics show a relatively small distance for log-linear, gamma, and the exponentials.

Tab. 8.2: Distribution of pause duration in older and younger speakers of English and Korean. Older speakers' pauses are shorter on average, and the differences between the cohorts are not significant in both languages. Pause production seems to be a stable aspect of speech production. The pause-to-word ratio increases in older speakers, indicating that the relative pause duration decreases as a function of the increasing utterance length in older speakers.

	Pause Duration					
Speaker	n	mean(SD)	Mood Z	p	Mann-Whitney U	p
ENGLISH	40	404.55 (118.40)				
Younger	20	408.21 (113.43)				
Older	20	400.88 (126.03)				
			-0.95	0.3398	0.1	0.5291
KOREAN	40	447.12 (124.43)				
Younger	20	454.96 (128.90)				
Older	20	439.28 (122.62)				
			-0.12	0.9081	0.0698	0.6588
Aggregate	n	mean(SD)	word-pause-ratio (p/w)	R^2	slope	
ENGLISH	26952	412.14 (426.53)	10.77 (0.928)	0.982	-0.102	
Younger	13242	413.00 (430.02)	10.16 (0.098)	0.982	-0.102	
Older	13710	411.32 (423.16)	11.37 (0.088)	0.983	-0.104	
KOREAN	21572	463.88 (378.18)	10.74 (0.931)	0.996	-0.119	
Younger	10259	475.55 (393.58)	9.99 (0.100)	0.997	-0.115	
Older	11313	453.30 (363.35)	11.42 (0.088)	0.993	-0.123	

Tab. 8.3: Distribution of utterance position in older and younger speakers of English and Korean. Older speakers' produce longer utterances on average, and the differences between the cohorts' means are significant in both languages. The variability of utterance length decreases in Korean and increases in English. Accordingly, the difference in cohorts' median is significant in the English, but not in the Korean sample. Remarkably, the individual variability increases in older speakers of English and decreases in individual speakers of Korean. By contrast, standard deviation increases in both languages' aggregates, indicating that the variability is distributed across contexts (utterances) in Korean, while in English the variability increases across contexts and individuals.

Utterance Position						
Speaker	n	mean(SD)	Mood Z	p	Mann-Whitney U	p
ENGLISH	40	5.82 (1.13)				
Younger	20	5.50 (0.67)				
Older	20	6.13 (1.40)				
			2.43	0.01	0.28	0.0763
KOREAN	40	4.06 (0.85)				
Younger	20	3.74 (0.80)				
Older	20	4.37 (0.78)				
			-1.13	0.2573	0.47	0.0032
Aggregate	n_{utt}	n_{words}	mean(SD)	R^2[3]	slope	
ENGLISH	118430	290318	5.710 (5.36)	0.999	-0.203	
Younger	57978	134479	5.446 (5.07)	0.995	-0.222	
Older	60452	155839	5.960 (5.61)	0.999	-0.190	
KOREAN	128317	231632	4.017 (3.30)	0.999	-0.322	
Younger	63175	102451	3.701 (3.01)	0.999	-0.358	
Older	65142	129181	4.309 (3.52)	0.998	-0.304	

[3]Fit to geometric distribution.

Tab. 8.4: Difference between consecutive pauses (nPVI) as a smooth function of pause duration and speaker age.

ENGLISH				
A. parametric coefficients	Estimate	Std. Error	t-value	p-value
Intercept	0.4297	0.0015	278.0719	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
s(pause):older	3.9767	3.9996	187.4378	< 0.0001
s(pause):younger	3.9762	3.9996	247.2637	< 0.0001
KOREAN				
A. parametric coefficients	Estimate	Std. Error	t-value	p-value
Intercept	0.3462	0.0015	235.0570	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
s(pause):older	3.9752	3.9996	251.9729	< 0.0001
s(pause):younger	3.9755	3.9996	217.5087	< 0.0001

Tab. 8.5: Mean pause duration as a function of utterance length and age, with a factor smooth for phrase boundary (final/non-final)

ENGLISH				
A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	6.0247	0.0080	756.2370	< 0.0001
phrase final	-0.0092	0.0196	-0.4697	0.6386
B. smooth terms	edf	Ref.df	F-value	p-value
s(utterance length):older	1.0003	1.0006	6.5620	0.0104
s(utterance length):younger	1.0002	1.0004	0.0309	0.8610
KOREAN				
A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	6.0902	0.0069	879.6565	< 0.0001
phrase final	0.0598	0.0160	3.7275	0.0002
B. smooth terms	edf	Ref.df	F-value	p-value
s(utterance length):older	1.0004	1.0009	1.0288	0.3105
s(utterance length):younger	1.0003	1.0006	0.4286	0.5128

8.1 Chapter 7, Supplements

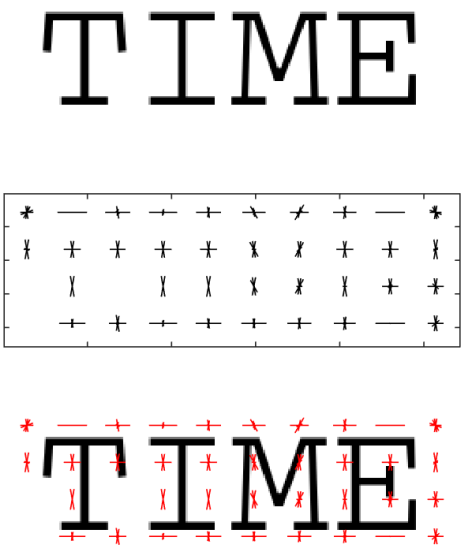

Fig. 8.3: Histogram of oriented gradient features extracted for stimuli