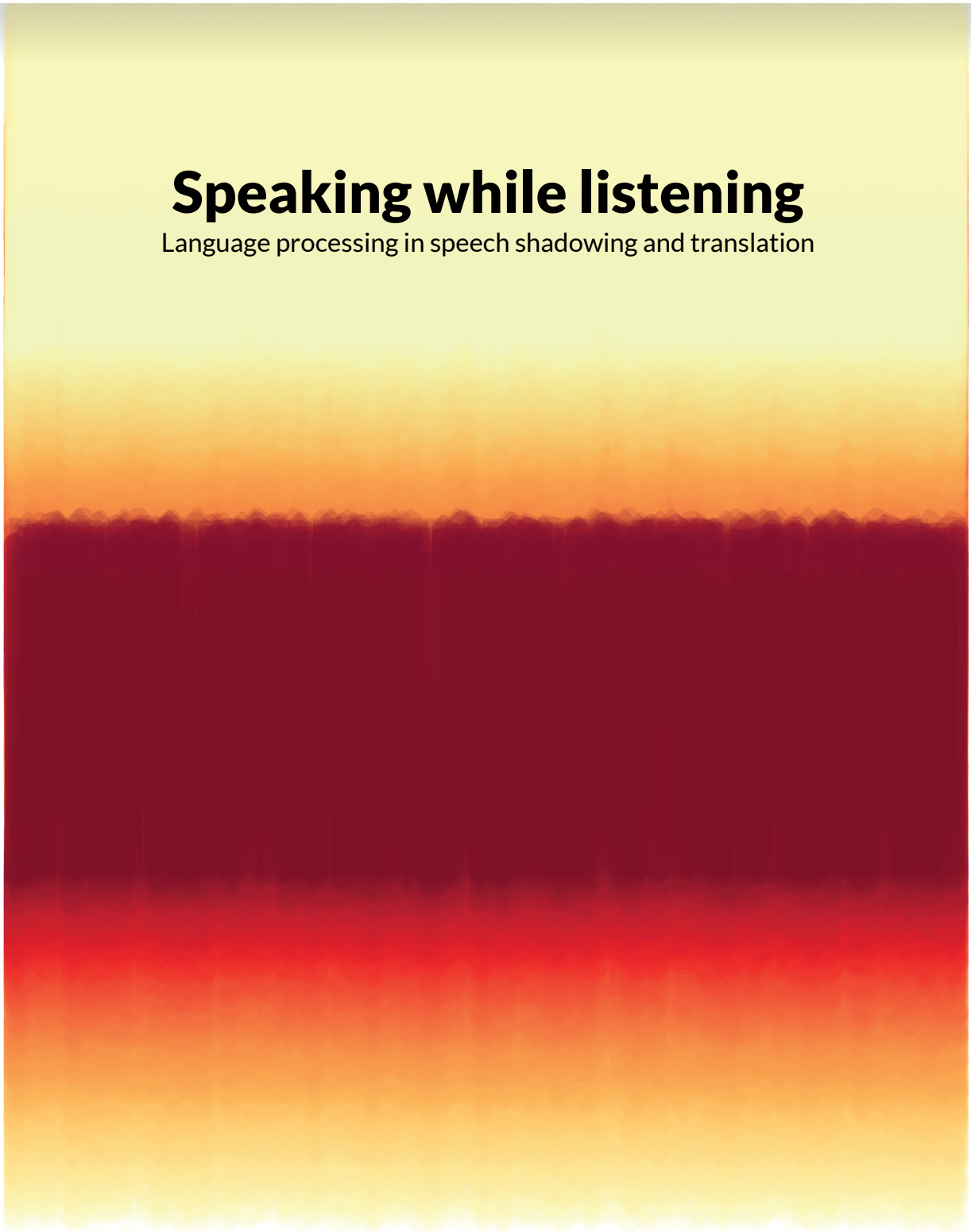
MAX PLANCK INSTITUTE
FOR PSYCHOLINGUISTICS

Speaking while listening

Language processing in speech shadowing and translation



Jeroen van Paridon

Speaking while listening

Language processing in speech shadowing and translation

The educational component of the doctoral training was provided by the International Max Planck Research School (IMPRS) for Language Sciences. The graduate school is a joint initiative between the Max Planck Institute for Psycholinguistics and two partner institutes at Radboud University – the Centre for Language Studies, and the Donders Institute for Brain, Cognition and Behaviour. The IMPRS curriculum, which is funded by the Max Planck Society for the Advancement of Science, ensures that each member receives interdisciplinary training in the language sciences and develops a well-rounded skill set in preparation for fulfilling careers in academia and beyond. More information can be found at www.mpi.nl/imprs.

This research was conducted at the Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands with support from the Max Planck Society for the Advancement of Science, Munich, Germany.

The art on the cover of this book is meant to evoke a spectrogram, but it also alludes to the work of abstract expressionists like Barnett Newman, Helen Frankenthaler, Agnes Martin and in particular Mark Rothko. However, it is neither a spectrogram nor a painting, but procedural art; the brush strokes determined by word vectors from the datasets created in Chapter 6, for words taken from one of the stories used in Chapters 2 and 3. Code for reproducing the cover art (or producing your own) can be found at github.com/jvparidon/rothcoverart.

ISBN: 978-94-92910-28-8
Printed and bound by Ipskamp Drukkers b.v.

Speaking while listening

Language processing in speech shadowing and translation

Proefschrift

ter verkrijging van de graad van doctor

aan de Radboud Universiteit Nijmegen

op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,

volgens besluit van het college van de decanen

in het openbaar te verdedigen op dinsdag 25 mei 2021

om 16:30 uur precies

door

Jeroen Pieter van Paridon

geboren op 28 augustus 1990

te Voorburg

Promotoren:

Prof. dr. Antje Meyer

Prof. dr. Ardi Roelofs

Manuscriptcommissie:

Prof. dr. Wilbert Spooren

Dr. Sybrine Bultena

Prof. dr. Rob Hartsuiker (Universiteit Gent, België)

This is really more of a comment than a question—

Contents

1 | Introduction

In settings where people who speak different languages need to understand each other, particularly when misunderstandings could cause serious problems, interpreters are relied upon to provide fast and accurate translations. Often interpreting is done consecutively, where the interpreter waits for a speaker to produce a short message before translating it for the intended recipient. In certain settings the delay in translation that is inherent to consecutive interpreting can be undesirable, for example if the speaker is making a speech for a large group of people, or if a speaker is testifying during a trial. In such scenarios, interpreters rely on simultaneous interpreting, a mode of interpreting where sentences are translated online almost as soon as they are uttered, leaving only as much latency as necessary for correct comprehension and translation of the speaker's message.

Simultaneous interpreting is a highly skilled profession that generally requires a postgraduate degree and takes years of training to fully master. Simultaneously attending to a speaker's message and formulating and producing a translation requires careful allocation of attention and a high degree of fluency in both the source and the target language. Language pairs can differ in word order, making it necessary to keep whole phrases or sentences in working memory in order to correctly reformulate them into the target language word order. Even when language pairs are closely related, interpreting can offer up unexpected challenges, for example in the form of "false friends" (cognates that are not literal translations).

While interpreting has a long history (with documented use at various royal courts going back many centuries, for instance), simultaneous interpreting specifically is a surprisingly recent phenomenon. The first documented use of simultaneous interpreting in a public forum was at the Nuremberg trials in 1945-1946 (Gaiba, 1998). For the first few years, there was no serious study of the mechanisms underlying simultaneous interpreting, with research focusing on issues with more direct practical consequences, such as the ethics of interpreting and the quality of the produced translations. Starting in the late 1960s however, several researchers Barik (1973), Gerver (1969, 1974a, 1974b, 1975), and Goldman-Eisler (1972) started performing experiments with interpreters, systematically manipulating the input stream in various ways and examining the quality of the produced output stream. Other researchers joined in, proposing theoretical models of the cognitive mechanisms underlying simultaneous interpreting (Gerver, 1975; Moser, 1978).

In that same era, Marslen-Wilson (1973, 1975) started researching speech shadowing, another paradigm that requires concurrent speech perception and speech production. In speech shadowing, participants are instructed to repeat a spoken message as quickly as they can manage, but in the original language, rather than in another language. In this paradigm there is no need to wait for a coherent message before initiating translation, resulting in very short latencies (at least in some participants, for more on the difference between "close" and "distant" shadowers see Marslen-Wilson, 1985). At such short latencies, speech is often reproduced before comprehension is fully realized, with participants in effect bypassing their conceptual processing to some degree in order to achieve the shortest possible latency. This is a remarkable adaptive mechanism that enables much faster reproduction, but in simultaneous interpreting, such a mechanism is not feasible. Listening to one language and reproducing the message in another requires comprehension to reach the conceptual level before production can be

initiated. Even if we allow for interlingual links at the lexical level, complete translation of a message would require a complete conceptual representation, especially in the case of idiomatic expressions or interpreting between languages with large differences word order.

Most of the recent research into simultaneous interpreting does not concern itself with these low-level psycholinguistic processes, but is instead focused on cognitive differences between interpreters and regular bilinguals or monolinguals, driven by the idea that mastering a skill as cognitively demanding as interpreting might convey more general cognitive benefits (e.g., improved working memory capacity). Many of these studies find that interpreters perform better than normal bilinguals on a variety of working memory tasks (see e.g., Stavrakaki et al., 2012), semantic fluency tasks (see e.g., Stavrakaki et al., 2012), a Simon task (Woumans et al., 2015; but for contradictory findings see Yudes et al., 2011), the Attentional Network Task (Woumans et al., 2015), and the Wisconsin Card Sorting Task (Yudes et al., 2011). An alternative explanation, that interpreters (and foreign language teachers) might be (self-)selected for these professions based on these cognitive traits, seems unlikely given the finding that trainee interpreters are intermediate between normal bilinguals and professional interpreters in working memory performance, and continue to improve during training. Further evidence for the cognitive differences resulting from adaptation rather than innate ability comes from a longitudinal fMRI study showing structural and functional adaptation over the course of a year-long training programme (Hervais-Adelman, Moser-Mercer, & Golestani, 2015).

A general caveat to this line of interpreting research is that cognitive advantage studies which include a professionally bilingual control group (such as teachers of a foreign language) generally find no significant difference between interpreters and professional bilinguals (see e.g., Christoffels et al., 2006), which suggests that any cognitive advantage might be related to better

second language proficiency, frequent code-switching, or frequently performing translation-related tasks. However, this notion of a benefit to general cognitive ability for highly proficient bilinguals is itself controversial, with several recent meta-analyses of bilingual cognitive advantage studies failing to find evidence in support of the notion that bilingualism conveys benefits in terms of cognitive abilities such as task-switching and working memory capacity (see e.g., Lehtonen et al., 2018; Paap et al., 2019). Most recently, Santilli et al. (2019) compared professional interpreters and proficient bilinguals specifically on language-related tasks, rather than the usual cognitive control and working memory measures. Interpreters scored better than bilinguals on interpreting-specific skills such as word translation, but not on more general language tasks such as picture naming and word reading, suggesting that interpreting experience benefits a very narrow set of skills, rather than a broad set of executive functions. Whether a so-called "interpreter advantage" exists in any meaningful sense therefore remains an open question, one that will surely drive further interpreting research for years to come. In this dissertation, however, I focused not on the supposed cognitive benefits of learning to perform simultaneous interpreting, but on the act of simultaneous interpreting itself, how it relates to speech shadowing, and its implications for our understanding of speaking and listening more generally.

Speech production and speech comprehension, although linked by the simple fact that the output of the former is the input for the latter, are most often studied in isolation from each other. In experiments investigating the mechanisms of speech comprehension, the participant is often tasked with producing a non-verbal and non-linguistic (or verbal but only nominally linguistic, for instance "yes" or "no") response to linguistic stimuli, while in speech production research, the participant is tasked with producing a spoken response to non-linguistic stimuli or linguistic stimuli presented in a very constrained fashion. The interaction and overlap between production and comprehension

mechanisms is thus fairly poorly understood, despite evidence that temporal overlap between comprehension and production processes is quite common. In conversational turn-taking, for instance, there is a very short (and sometimes non-existent) latency between turns, often cited as evidence for the idea that the second speaker is able to plan at least part of their speech in advance while listening to (and comprehending) the speech of the first speaker (see Levinson, 2016, for review).

Early research into interaction between production and comprehension was focused on the intelligibility of speech presented to a participant in the midst of producing unrelated speech. While such studies were not driven by any detailed models of production and comprehension and therefore not designed to test detailed predictions about production or comprehension, these studies nevertheless started a fruitful tradition. Arguably one of the most useful insights about the interaction between production and comprehension comes from Broadbent (1952) who, in an early study of simultaneous production and comprehension, states about the difficulty of the task: "One aspect of this problem is the relation of the behavioral mechanisms used in speaking to those required for listening, since both may be needed simultaneously" (p. 267).

Barik (1973) claims that interpreters partly get around this limitation by making use of the naturally occurring pauses in the input speech for their output production, but Gerver (1975) notes that this cannot be sufficient time for the interpreter articulate a full translation of the input speech. More specifically, if output were produced in the pauses in the input, pauses would need to make up at least half of the input speech signal (assuming similar speech rates for input and output speech), but Gerver (1975) demonstrates empirically that the input speech contains fewer pauses than would be required and that the input speech and output speech instead overlap significantly. Goldman-Eisler (1972) goes further yet in claiming that interpreters largely ignore the chunking and pauses in

the input speech, noting that 90% of interpreter utterances are either composed of multiple input chunks or initiated before an input chunk has ended. Pauses are demonstrably used in some strategic fashion by interpreters, as evidenced by the comparatively larger portion of output speech articulated during pauses than during input speech, but as Gerver (1975) states, this does not necessarily mean that pauses are essential to the process. Regardless of how pauses are used, it is clear that to the extent that speech production and comprehension occur concurrently in simultaneous interpreting, a degree of independence between comprehension and production processes is required.

Christoffels and De Groot (2004) postulate the existence of separate input and output lexicons, as well as separate lexicons for each language, enabling the selective inhibition of the L2 output and L1 input lexicons during interpreting. This is an effective way of solving the problem of simultaneous lexical processing for both comprehension and production, to be sure, but it is not very plausible. One only needs to consider the problems language acquisition would run into if input and output lexicons were entirely separate: hearing a word used in context would not be sufficient to learn how to produce the word, instead, production of every word would need to be learned by trial and error, despite knowledge of the word already being present in the comprehension domain. While early acquisition may involve plenty of trial and error learning, acquisition of new words in older children and adults is trivial: comprehension of a word in its proper context enables use of that word in production from that point onward (as long as it is recalled). For example: if I buy a new kind of fruit at the grocery store, and the cashier tells me it is called a "torp", I will be able to tell anyone I meet from then on that I have bought (and eaten) a "torp". The word will be available for production without any conscious transfer. Ultimately, the question of how comprehension and production processes are temporally coordinated during simultaneous interpreting (and speech shadowing) so as not to interfere with each

other remains open. The aim of this dissertation is to put forth a more plausible explanation and provide experimental evidence for it.

Existing process models of interpreting are generally box-and-arrow models without a computational implementation. These models are generally useful graphical representations of the chain of processes required to successfully perform simultaneous interpreting, but as a means of testing theories about cognitive processes they fall short because they do not make falsifiable predictions about quantitative aspects of behavior (e.g., speech latencies, pause durations, or error rates). Similarly, purely descriptive statistical models of experimental data can confirm quantitative behavioral predictions in an experimental paradigm, but a computationally implemented process model is essential because it would allow us to generate such quantitative predictions.

In my view, a satisfactory process model of temporal coordination of simultaneous speech production and comprehension (as in simultaneous interpreting and speech shadowing) should allow for at least partially overlapping production and comprehension, be broadly consistent with the existing psycholinguistic literature (concerning e.g., response selection processes, the time course of lexical access in comprehension and production, etc.) and explain the *seemingly* concurrent use of shared resources and mechanisms without resorting to implausible theories (e.g., the model should not require separate lexicons for production and comprehension). To validate such a process model it should be implemented computationally, and its predictions tested against behavioral data such as error rates or speech latencies from simultaneous interpreting and speech shadowing.

1.1 Overview of the upcoming chapters

In Chapter 2, I introduce a process model of speech comprehension and production in speech shadowing and simultaneous interpreting. The processing stages and starting values for the durations of those stages are derived from a review of the psycholinguistic literature, but I find that these literature-derived parameters cannot fit the error rates I observe in behavioral experiment unless I introduce a novel mechanism or parameter. I hypothesize a switch cost is incurred each time interpreters switch between comprehension and production at the lexical processing stage. If this switch cost is set at approximately 50 milliseconds, our model closely fits the observed error rates.

In Chapter 3, I collect novel data to further test the model proposed in Chapter 2. I find that performance on randomized word lists (i.e. text without semantic or syntactic structure) is not consistent with model predictions. Varying the durations of processing stages in the model within psycholinguistically plausible bounds also does not result in an acceptable fit. I take this as evidence that my model, as originally proposed, lacks a top-down contextual facilitation component that could plausibly explain the difference in performance on narrative versus non-narrative text.

In Chapter 4, I attempt to uncover the locus and magnitude of this top-down contextual facilitation. I pool the data from Chapters 2 and 3 and reannotate them for speech latencies; I also compile a wide variety of lexical and contextual covariates for this dataset. I then use a multilevel Bayesian regression to examine the relative effects of these lexical and contextual factors on speech latencies in speech shadowing and simultaneous interpreting.

Chapters 5 and 6 detail research conducted in service of the study reported in Chapter 4. These chapters do not deal with simultaneous language comprehen-

sion and production directly, but concern psycholinguistic research methods in a broader sense.

In Chapter 5, I expand on counterintuitive effects of transitional probabilities identified in Chapter 4, and demonstrate how these can arise from collinearities in linear models. I explore how frequencies and transitional probabilities are distributed in a representative corpus, and make a case for informed a priori selection between predictors known to be collinear.

In Chapter 6, I discuss the construction and validity of word embeddings, a class of models of distributional semantics (as used in Chapter 4) and introduce a novel set of word embeddings trained on subtitle corpora. I demonstrate that models trained on pseudoconversational text perform as well as, and in some cases better than, models trained on the most commonly used training corpora (Wikipedia text). A combined corpus outperforms both subtitles and Wikipedia, highlighting a need for contextual diversity in NLP training corpora.

In Chapter 7, I summarise the key findings from the preceding chapters and discuss how these findings relate to our broader understanding of comprehension and production processes as they occur during simultaneous interpreting and speech shadowing.

2 | A lexical bottleneck in shadowing and translating of narratives[1]

Abstract

In simultaneous interpreting, speech comprehension and production processes have to be coordinated in close temporal proximity. To examine the coordination, Dutch-English bilingual participants were presented with narrative fragments recorded in English at speech rates varying from 100 to 200 words per minute and they were asked to translate the fragments into Dutch (interpreting) or repeat them in English (shadowing). Interpreting yielded more errors than shadowing at every speech rate, and increasing speech rate had a stronger negative effect on interpreting than on shadowing. To understand the differential effect of speech rate, a computational model was created of sub-lexical and lexical processes in comprehension and production. Computer simulations revealed that the empirical findings could be captured by assuming a bottleneck preventing simultaneous lexical selection in production and comprehension. To conclude, our empirical and modelling results suggest the existence of a lexical bottleneck that limits the translation of narratives at high speed.

[1] Adapted from Van Paridon, J., Roelofs, A., & Meyer, A. S. (2019). A lexical bottleneck in shadowing and translating of narratives. *Language, Cognition and Neuroscience, 34*(6), 803–812. https://doi.org/10.1080/23273798.2019.1591470

2.1 Introduction

Simultaneous interpreting, also known as conference interpreting, is the online oral translation of spoken language. Most often used at international conferences and institutions, this mode of translation provides a near instantaneous translation to the listener. While there is an extensive (and often contradictory) literature on cognitive differences between interpreters and non-interpreter bilinguals (e.g., Morales et al., 2015; Woumans et al., 2015), the processes of speech comprehension and production occurring during simultaneous interpreting have not been studied in much detail. Behavioural studies have examined the linguistic skills involved in interpreting (Christoffels et al., 2003; Christoffels et al., 2006), and neuroimaging has started to identify the neural bases of interpreting (Hervais-Adelman, Moser-Mercer, & Golestani, 2015; Hervais-Adelman, Moser-Mercer, Michel, et al., 2015). However, no theory of interpreting exists that describes the time course of concurrent speech comprehension and production in simultaneous interpreting.

Even though professional interpreters are highly trained at concurrent listening and speaking, comprehension and production are still somewhat impaired by their temporal overlap during interpreting. More errors are made during interpreting than during simple shadowing, and interpreting a speech leads to significantly worse recall than simply listening to that speech (Gerver, 1974b). Additionally, interpreters cannot interpret at very high speech rates. In a seminal study, Gerver (1969) had six professional interpreters shadow recordings of diplomatic speeches played at different speeds, while six others interpreted the same recordings. For the materials used by Gerver, the maximum input speech rate for fluent French to English interpreting (with more than 90% of words being translated correctly) was around 110 words per minute on average, with performance declining linearly at higher input speech rates to less than 60% correct at 164 words per

minute. Below the maximum input rate, interpreters can approximately match the output speech rate to the input speech rate, producing mostly complete and correct translations. At higher input rates, interpreters start to omit words and phrases, and produce in short, high speech rate bursts. The maximum interpreting rate lies well below the maximal speech rate interpreters can comprehend or produce when not interpreting, as evidenced by Gerver's shadowers, who were still fluent at 142 words per minute. These differences suggest that the limit on interpreting rate is not set directly by limits on the processes of speech comprehension or production separately, but rather by limits on the speech system as a whole arising when comprehending and producing speech concurrently.

How this coordination is achieved in a fluent manner, why it breaks down at high input speech rates, and how the resulting error pattern comes about is not explained by any of the relatively few models of interpreting that have been put forward in over half a century of interpreting research. Models of simultaneous interpreting can be grouped into several categories. One type is the effort model proposed by Gile et al. (1997), which poses that interpreting consists of four different types of effort: listening, production, memory, and coordination. These types of effort are assumed to be additive and to simultaneously require capacity. It is not apparent, however, how this model might be tested empirically. Another type of model is the process model that describes the organisation of processing in interpreting. This type of model tends to resemble a complex flowchart of processing steps, but none of the existing models makes falsifiable predictions about measurable indices of interpreting processes such as timing, error rates, or error types (Gerver, 1975; Mizuno, 2005; Moser, 1978). The only interpreting model that has been empirically tested is the cognitive load model by Seeber and Kerzel (2012), which makes predictions about physiological indices of cognitive load (i.e. pupil diameter) based on hypotheses about the processing demands of different types of linguistic input. Seeber and Kerzel found that translating German

SOV (subject-object-verb) sentences into English SVO (subject-verb-object) sentences produced a marginally higher cognitive load than translating from SVO into SVO sentences. Their examples of SOV sentences include long-distance dependencies, however, which could explain the increased cognitive load regardless of task demands specific to interpreting. Despite this apparent confound, their model suggests that word order might play a role in interpreting performance, but does not explain the specific limits on interpreting speech rate and the associated error patterns.

A model of simultaneous interpreting of the type that Gerver (1975) suggested, that is a model that explains all of the linguistic and metalinguistic processes a professional interpreter relies on, cannot currently be specified in quantitative terms such as latencies or error rates. This is because we do not have a sufficiently detailed understanding of all the processes involved. However, a simpler, purely lexical model that explains only the simultaneous word comprehension and production aspect of interpreting can be generated by combining behavioural data of the type collected by Gerver (1969) with current psycholinguistic models of word production and comprehension. Such a model would not be a complete model of interpreting but it could extend experimentally supported psycholinguistic models of word production and comprehension and describe the coordination of production and comprehension, which is one of the key elements of simultaneous interpreting. Showing that such a model predicts error rates in simultaneous production and comprehension would demonstrate that interpreting is subserved by normal language processing, albeit under abnormal task demands. More generally, the specific adaptations needed to simulate the error rates reported by Gerver (1969) could provide new insights into the way comprehension and production are coordinated when fluid and frequent transition between the two is required.

Prior work on spoken word production in a dual-task paradigm has demonstrated that semantic interference in a production task can cause delays in response selection for a second, unrelated task performed at the same time, whereas a phonological effect in the same production task does not always propagate to the second task. This suggests that central attention is required for response selection at the lemma level, but not (or less) at the phonological level (Cook & Meyer, 2008; Ferreira & Pashler, 2002; Piai et al., 2014; Roelofs, 2008; Roelofs & Piai, 2011). Having to coordinate selections at the lemma level for both comprehension and production could conceivably create a lexical-selection bottleneck during interpreting. The present study examined whether a computational model of interpreting and shadowing that includes such a bottleneck could account for error rates in relevant behavioural data. This was done by adding a lexical bottleneck to the model of word production and comprehension proposed by Indefrey and Levelt (2004).

Of course, generally speaking, syntactic processing must be an important component of interpreting, especially where the source text and the correct translation differ in word order. However, in our texts the English and Dutch word order were mostly the same. Moreover, the Indefrey and Levelt (2004) model does not describe syntactic processes beyond the assumption that lemma selection affords access to the syntactic properties of a given word. Therefore, we chose, as a parsimonious starting point for the model, not to include an explicit processing cost for syntactic processing, but rather to test whether the lexical model suffices to simulate the relevant behavioural data.

To be able to test the model empirically, we first collected relevant behavioural data. To this end, we repeated Gerver's (1969) study comparing interpreting and shadowing performance at different speech rates, but with a more rigorously controlled design. The languages involved in the present study were English and Dutch. Our design was a within-participants comparison of shadowing and inter-

preting performance with source texts presented at a range of speech rates. We recruited participants without prior interpreting training, to exclude the possible use of interpreting-specific processing strategies. The behavioural data were then used to fit our computational model. Note that the present work is concerned with switching between production and comprehension (either within the same or in different languages) as required in the shadowing and translation tasks, and not with the (code) switching between L1 and L2 speech production, which is often considered in studies of bilingualism. Switching between comprehension and production in interpreting is a type of switching between L2 and L1 that is not required during shadowing. But because in interpreting L2 is used exclusively for comprehension and L1 for production, there is no need for language-specific inhibition of response-selection in production, which is often hypothesised to be the cause of bilingual switch costs (e.g., Meuter & Allport, 1999).

2.2 Methods

2.2.1 Participants

The participants were native speakers of Dutch, recruited from the participant pool at the Max Planck Institute for Psycholinguistics. To identify Dutch-English bilinguals with sufficient proficiency in English to perform the tasks, 215 participants were screened using LexTale, an online English vocabulary test, which correlates well with other measures of English proficiency (Lemhöfer & Broersma, 2012). From this group, participants with a LexTale score over 85% (the top 33% of test takers) were invited to participate in the study. Of the invitees, 20 agreed to take part in the study (13 female, mean age 22.3 years). Mean self-reported age of acquisition of English was 10.3 years (SD = 1.1 years, N = 14), approximately the age at which English education starts in Dutch primary schools. None of the par-

ticipants had prior experience in shadowing or interpreting; their mean LexTale score was 91.4% (SD = 5.1%).

2.2.2 Procedure and design

Participants performed two sessions of shadowing and interpreting, one week apart. Both sessions were recorded but the first session was meant solely to familiarise participants with the tasks and was not analysed. The second session consisted of two blocks of roughly 20min: one shadowing block of five spoken texts presented at different speech rates and one interpreting block of five spoken texts presented at different speech rates. The order of texts, tasks, and speech rates was counterbalanced across participants so that each text was presented to two participants at every speech rate and in both tasks, but texts and speech rates were not repeated within participant.

2.2.3 Materials

Stimuli were ten samples of around 300 words in length, taken from a variety of books for children between six and ten years. Children's books were selected because they feature few rare lexical items and few complex syntactic structures that would require extensive reformulation during translation. Using a teleprompter script, the sample texts were recorded by a male native speaker of English at a controlled rate of 150 words per minute. To produce the desired stimulus speech rates, the recordings were sped up or slowed down to 100, 125, 150, 175, and 200 words per minute using the Audacity audio editor (Version 2.0.6, Audacity Team, 2017). Because the digital speech rate manipulation produces audible distortions in the recordings, the stimulus texts were rerecorded by the same speaker while playing the digitally sped-up or slowed-down recordings over headphones as a continuous speech rate cue.

2.2.4 Analysis

Shadowing performance was scored by transcribing participant recordings and counting the percentage of words correctly reproduced from the source text. Interpreting performance could not be scored so straightforwardly; instead, recordings were transcribed by native speakers of Dutch and the percentage of words from the source text that was represented in the transcription was taken as the score. Scoring was double-checked by a second native speaker of Dutch.

Speech rate, task, and interaction effects on performance were analysed with a logistic mixed-effects model using the lme4 package (Bates et al., 2015) in R (Version 3.3.3, R Core Team, 2013). Statistical inference for the coefficients was computed using the lmerTest package (Kuznetsova et al., 2017).

2.3 Results

Figure 2.1 shows participants performance. As expected, both shadowing and interpreting performance decreased as the source text speech rate increased. However, the difference in slopes of interpreting (.26% per wpm) and shadowing (.17% per wpm) performance across source speech rates indicates an interaction effect.

When controlling for random effects of participant and source text with random intercepts, there were significant effects of task ($\beta = 0.63, SE = 0.014, p < .001$), source speech rate ($\beta = 0.62, SE = 0.014, p < .001$), and the interaction between task and source speech rate ($\beta = 0.05, SE = 0.014, p < .001$). Controlling for random effects of participant and source text using a more elaborate random effects structure with random slopes was not possible due to the limited number of observations in each cell.

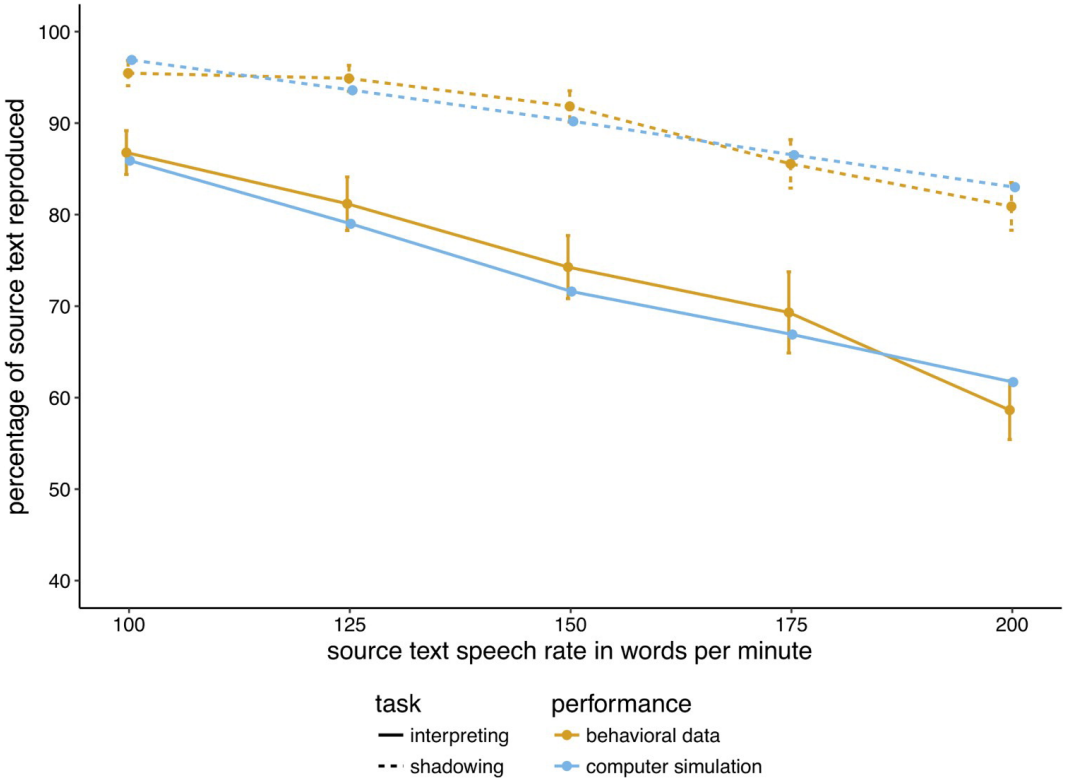


Figure 2.1: Mean percentage of source text reproduced while shadowing and interpreting across source speech rates in the behavioural experiment and model. Error bars represent standard error for the behavioural data.

2.3.1 Interim discussion

Shadowing performance was at ceiling at 100 and 125 words per minute, which is consistent with the shadowing performance reported by Gerver (1969). Aside from the ceiling effect, the decrease in performance was roughly linear for both interpreting and shadowing. While there appears to be an interaction of source speech rate and task in the data reported by Gerver similar to the interaction in the present study, the deterioration in interpreting performance with increasing source speech rate was more severe in Gerver's data (from 95% correct to less than 60% correct for a 70% increase in speech rate) than in the present study (from 87% correct to 59% correct for a 100% increase in speech rate). This difference may have been caused by the nature of the source materials used by Gerver, although this should be partially mitigated by the professional-level proficiency of the participants in that study. Another possible cause is the design used by Gerver: a between-participants design with only six participants performing each task. Such a design is likely to be underpowered and more susceptible to noise than the present within-participants design with 20 participants performing both tasks. Regardless of the quantitative differences between the present results and those reported by Gerver, the notion that there are different factors limiting interpreting and shadowing performance at high source speech rates is supported by the interaction of source speech rate and task in both studies.

What these factors that limit interpreting and shadowing are is not obvious from the behavioural data. Participants reported feeling that interpreting at high speech rates required alternately attending to input and producing output, as in task switching. If one of these tasks takes too long (e.g., a long sequence of words needs to be produced, but in the meantime new input is coming in) words are lost, often several at a time. At lower speech rates, participants reported that interpreting felt more natural or automatic, which either reflects an ability to gen-

uinely attend to both comprehension and production at the same time or more fluent task switching that participants are not as aware of.

One difference between the shadowing and interpreting tasks that potentially modulates the effect of task on performance is the production language. During interpreting the participants spoke in their native language (Dutch), while during shadowing they spoke in their second language (English). Participants were screened for English proficiency, but most likely production was easier in their native Dutch. However, any native language advantage would only serve to increase performance in the interpreting task and therefore decrease or mask the task effect.

2.4 Computational model

The observed interaction between task and speech rate suggests a temporal coordination problem that causes shadowing and interpreting performance to differentially degrade with increasing speech rate. To identify the source of that problem, we constructed a simple computational model. Based on dual-task studies of language production (Cook & Meyer, 2008; Ferreira & Pashler, 2002; Piai et al., 2014; Roelofs, 2008; Roelofs & Piai, 2011), we set out to test the assumption of a lexical-selection bottleneck. The model represents the combined model of speech comprehension and production presented by (Indefrey & Levelt, 2004), implemented as a chain of consecutive processing stages, as illustrated in Figure 2.2. It takes as input a sequence of words and their onset times. To replicate the behavioural paradigm as closely as possible, the recordings presented to the participants were used to generate these input sequences for the model. We used the WebMAUS automatic speech segmentation service to assign onset times to each word in an orthographic transcription of the recordings (Kisler et al., 2017).

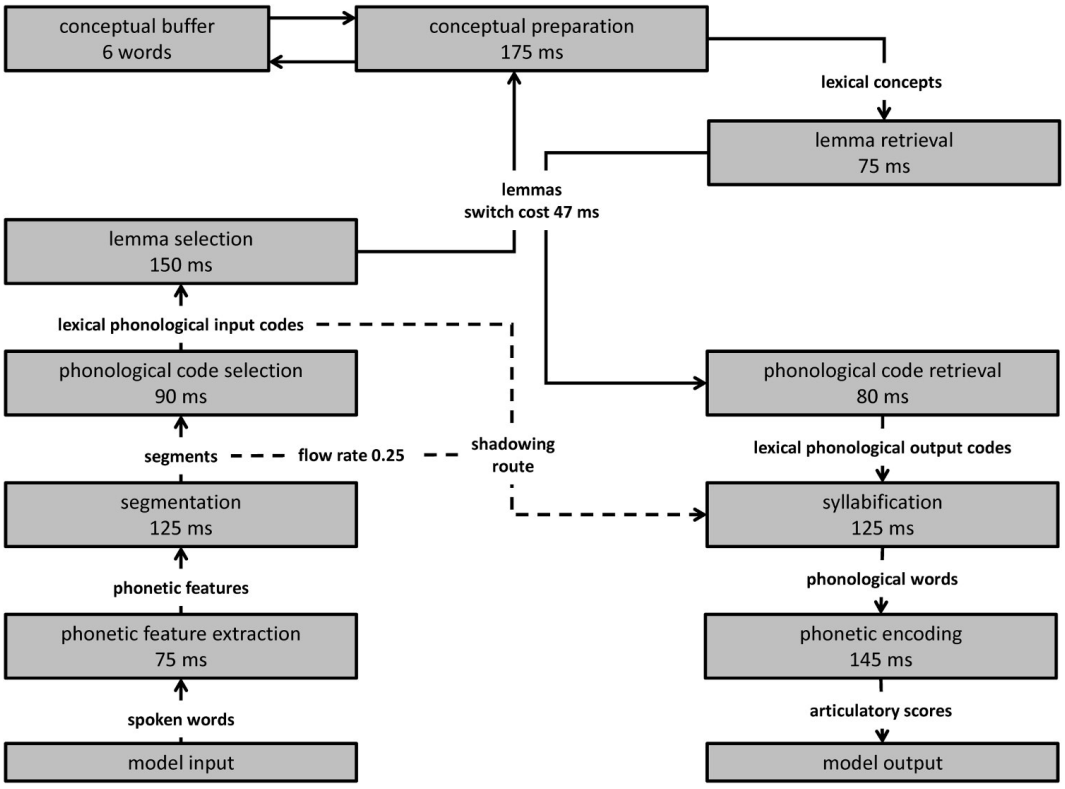


Figure 2.2: Structure and parameters of the computational model. The solid arrows are hypothesised to represent a route used in both simultaneous interpreting and shadowing, while in shadowing many words can also be reproduced along the route represented by the dashed arrow. Parameters set using particle swarm optimisation are conceptual buffer size, lemma switch cost, segmentation-syllabification accumulation rate, and function word accumulation rate.

Each word starts at the first processing stage in the model and is then passed along after being processed for a specific duration. This was implemented computationally by representing each processing stage as a simplified linear ballistic accumulator; the simplification being that the rate of evidence accumulation was fixed to a rate at which the time to reach threshold matches the durations reported by Indefrey and Levelt (2004) and Indefrey (2011) instead of drawing the accumulation rates from a normal distribution as originally proposed by Brown and Heathcote (2008). This simulated processing does not comprise any sort of linguistic processing because the model operates only on the onset times of the words. Details of component processes were unimportant as only the latencies of the processes and their interdependencies mattered (cf. Schweickert, 1980). Professional interpreters likely use interpreting-specific strategies to facilitate processing, but because we were attempting to model the error rates of untrained interpreters we did not attempt to model these processing strategies. Modelling interpreting-specific strategies might also reduce the validity of the model for describing the coordination of language comprehension and production in contexts other than interpreting.

To account for the reduced processing demands of function words compared with content words, function words were assigned an increased rate of evidence accumulation. As an initial estimate, the rate of evidence accumulation was set to double that of content words, but this value was later adjusted in a parameter optimisation procedure described below. On average, 52.8% of the words in a stimulus text were classified as function words. For a complete list of the words classified as function words in the present study see Appendix A in the Supplemental Materials.

The duration of conceptual processing was difficult to derive, as Indefrey and Levelt (2004) based their estimates on picture naming and single word listening, instead of a task that involves sentences and combines both speech comprehen-

sion and production. One commonly used experimental paradigm that requires sequential comprehension and production is single word translation. However, the reported latencies for single word translation vary from roughly 800 ms when words were presented orthographically (La Heij et al., 1996), to as much as 1200 ms when words were presented auditorily (De Groot, 1992). As an initial approximation, therefore, we adopted the 175 ms estimate reported by Indefrey and Levelt, because even though that estimate is derived from picture naming experiments, it leads to an overall single word interpreting latency that roughly matches the latencies reported by De Groot.

The remaining component process durations were also based on the latencies reported by Indefrey and Levelt (2004), and the resulting model fit was measured as root-mean-square deviation (RMSD) from mean participant performance across texts for each speech rate in the behavioural experiment. The initial model was a poor fit for the behavioural data (combined RMSD=5.4%). After observing that the poor fit was caused in part by the model systematically underperforming in the shadowing task, we added an extra connection from segmentation to syllabification to improve shadowing performance. The extra connection reflects the unique affordance in shadowing of starting selection of an output phoneme directly after identifying an input phoneme because the output is identical phoneme-for-phoneme to the input (cf. Roelofs, 2004, 2014). The existence of such a low-level connection is supported by the short latencies found in previous shadowing experiments (e.g., Fowler et al., 2003). The connection was implemented by allowing additional evidence accumulation for syllabification from the moment segmentation is completed. Setting the accumulation rate through this additional segmentation-syllabification connection to an initial value of .5 markedly improved the model fit (combined RMSD=2.6%).

To improve the fit of the model performance in the interpreting task, we first introduced a conceptual buffer into the model to make it more closely resemble human processing of consecutive words. This addition was based on the observation that it is not only possible to conceptually combine the meanings of a set of words and to reorder them before production, but that this is required during interpreting. In our computational model, we capture the function of the conceptual workspace by assuming a buffer that holds concepts until they can be passed to lemma selection for production.

Next, we implemented the critical assumption of a lexical-selection bottleneck (Cook & Meyer, 2008; Piai et al., 2014; Roelofs, 2008). The bottleneck was implemented at the lemma level, which is assumed to be shared between production and comprehension (e.g., Levelt et al., 1999; Roelofs, 2004, 2014). Lemma selection for production was blocked while selecting a lemma for comprehension, and vice versa. In translating, a switch is required between comprehension in one language (English) and production in another (Dutch), which results in a switch cost (see e.g., Monsell, 2003, 2015, for review). This switch cost means that delay of access to the lemmas from the production stream can last for multiple words if new input words come into the comprehension stream close enough together to not allow time to switch back to lemma selection for production in the meantime. To determine the optimal switch cost and conceptual buffer size we implemented a parameter optimisation procedure. Function word accumulation rate factor and segmentation-syllabification accumulation rate were also entered into the parameter optimisation procedure.

To optimise our simulation of participant behaviour, we minimised the model's RMSD from the mean participant performance across texts for each speech rate in the behavioural experiment for both tasks by varying its free parameters using particle swarm optimisation implemented in the Optunity parameter optimisation library (Claesen et al., 2014). Particle swarm optimisa-

tion uses a swarm of communicating particles moving through the parameter space looking for an optimal parameter set. Particle swarm optimisation does not use a gradient for optimisation and is therefore well suited to the problem of optimising the parameters of this model (Kennedy & Eberhart, 1995).

From 1920 iterations (96 particles for 20 generations), we selected the parameter set that produced the performance closest to that of the participants. Optimal parameters were a function word evidence accumulation factor of 2.0, a segmentation-syllabification accumulation rate of .25, a buffer length of 6 words, and a lexical selection switch cost of 47 ms (combined RMSD=1.9%). This parameter set, and the model's structure, is reported in Figure 2.2.

Figure 2.1 shows the mean percentage of source text reproduced while shadowing and interpreting across source speech rates in the best-fitting model. In the simulations, the interpreting and shadowing performance progressively degraded with increasing speech rate, which corresponds to the empirical data. This degradation was stronger for interpreting than for shadowing, as empirically observed. Thus, the computational modelling suggests that to account for the data, it suffices to assume a lexical-selection bottleneck that precludes concurrent selection of lemmas in comprehension and production, and an associated switch cost.

2.5 Discussion

In the present study, we first replicated and expanded Gerver's (1969) study of interpreting and shadowing performance at different speech rates. We then used these data to test a computational model of interpreting and shadowing. The model structure and parameters were derived from a meta-analysis of speech comprehension and production experiments (Indefrey & Levelt, 2004). Performance of the computational model on a combined interpreting and

shadowing measure most accurately simulated behavioural data when lemmas could not be selected concurrently for production and comprehension, creating a lexical-selection bottleneck with an associated switch cost. This switch cost is a possible explanation for the emergence of task-switching type speech patterns at high speech rates, while at low speech rates comprehension and production seem to be temporally overlapping. The model suggests that temporal overlap is possible for processes such as phonetic decoding/encoding, phonological encoding/decoding, segmentation, and syllabification. Only lemma selection for comprehension and production cannot happen concurrently due to a lexical-selection bottleneck. At low speech rates, the switch cost can be absorbed into the pauses between words and the redundant parts of words that come after the uniqueness point. Therefore it is not perceivable to a listener that parts of production and comprehension are happening consecutively instead of concurrently. At higher speech rates the pauses are shortened and can no longer absorb the switch cost which then becomes a bottleneck, causing the model (and the participants) to periodically miss input or forget output, making the task set switching audible in the form of alternating bursts of listening and speaking.

Modelling the processing stages of speech comprehension and production as simplified linear ballistic accumulators makes the model as a whole computationally feasible, but it also necessitates that each consecutive step is discrete. This may be sufficient to capture the contrast between shadowing and interpreting, but it is important to note that recent more detailed models of single word production and/or comprehension (e.g., Roelofs, 2014; Ueno et al., 2011; Walker & Hickok, 2016) feature connectionist components that more plausibly simulate phenomena such as interaction and competition in the speech system. Implementing a plausible connectionist model and fitting its parameters was not feasible for this study. Given the large number of parameters present in such a model

and the relatively few data points it would be fitted to, there is an inordinate risk of overfitting. The lack of interaction between lower-level processes likely causes the present computational model to not capture facilitation or interference effects of cognates and incidental temporal coincidence of phonologically or semantically related words in the production and comprehension streams. It is unclear whether the net effect of this simplification in the model causes an over- or underestimation of the error rate. However, while the model does not capture small facilitation and interference effects, its contribution is that it postulates a critical path for both simultaneous interpreting and shadowing and demonstrates the temporal consequences for performance in both tasks. Future developments of the model could integrate cognate status and other lexical factors to allow for more specific predictions such as latency at the word level that cannot be derived from the current model. Computational models like the recent Multilink model proposed by Dijkstra et al. (2019) present estimates of the effects of lexical factors such as semantic equivalence of possible translations and cognate status for single word translation; such estimates could be incorporated into an interpreting model as well.

As our model is mostly blind to linguistic content (with the exception of the distinction between function words and content words) and has no knowledge of syntax, any kind of temporal clustering of errors is simply due to the time course of the input and the structure of the model. Assuming that the bottleneck is situated at the lexical level appeared to be sufficient to explain the data. The fact that the model still replicates the error rates observed in participants who have syntactic knowledge, and use it to reformulate English sentences into Dutch sentences, is striking. The lack of need for a syntactic component in the model suggests that in the present study syntactic processing did not impose a significant time cost, possibly be due to the simplicity of the stimulus texts and the close correspondence in word order between the two languages. Syntactic processing

is thought to be largely incremental in nature, both in production (Konopka & Meyer, 2014; Levelt, 1989) and comprehension (Altmann & Mirkovi, 2009; Christiansen & Chater, 2016), and once the entire message of a phrase is conceptualised, reformulating the type of short, grammatically straightforward sentences found in children's books may be so trivial that it does not cause meaningful additional cognitive load or delay. The occasional differences in word order between English and Dutch might merely require some extra time during conceptual processing, reflected in the small increase in conceptual processing duration needed for optimal model fit, when compared to the values Indefrey and Levelt (2004) report for conceptual processing during picture naming. For syntactically more complex texts or structurally more different languages, these model components and parameter values may be insufficient to model the syntactic costs. Under certain conditions, such syntactic costs may even constitute another bottleneck. In the present study, however, the model suffices to demonstrate that one important bottleneck is located at the lexical level.

2.5.1 Conclusion

Simultaneous interpreting and shadowing performance progressively degrades with increasing speech rate. This degradation is stronger for interpreting than for shadowing. Computational modelling showed that to account for the data, it sufficed to assume a lexical-selection bottleneck that precludes concurrent selection of lemmas in comprehension and production and causes the associated switch costs.

3 | Role of narrative context in interpreting and shadowing: performance on word lists

Abstract

In online interpreting and shadowing of spoken narratives, performance accuracy decreases more with increasing speech rate for interpreting than for shadowing. In Chapter 2 showed that a lexical bottleneck model correctly simulates the difference in performance between tasks at different speech rates. However, the absence of syntactic or semantic processing costs in the model may be questioned. In the present study, we directly tested the lexical bottleneck hypothesis by presenting the content words of narratives as randomised word lists to participants, and examined their shadowing and interpreting performance. We replicated the stronger deterioration of performance for interpreting than for shadowing with increasing speech rate. The deterioration for lists was even stronger than previously observed for narratives, which suggests that semantic and syntactic processing costs are entirely absent. This suggestion was corroborated by computer simulations, which revealed that our lexical bottleneck model likely underestimated the lexical switch cost in our previous study. We conclude that semantic and syntactic structure yield no costs but help in interpreting and shadowing narratives.

3.1 Introduction

An important finding on online translation of speech from one language into another, called interpreting, is that performance accuracy deteriorates with speech rate. In a seminal study, Gerver (1969) found a stronger deterioration of performance with increasing speech rate for professional interpreters performing an interpreting task than for interpreters performing a shadowing task (i.e., the literal repetition of speech). In Chapter 2 we replicated this finding for university students and simple spoken narratives. Moreover, we advanced a computational model that correctly simulated the behavioural error rates in both tasks.

The model used a parsimonious approach, starting with processing durations taken from Indefrey and Levelt (2004) and incrementally adding modifications to account for word-level interpreting accuracy. We accounted for the difference in performance accuracy between translating and shadowing at different speech rates by the inclusion of a lexical bottleneck. The bottleneck concerned a necessity to switch between lexical selection for comprehension and production with an associated switch cost of approximately 50 ms. Because the model allowed shadowing to be performed using a direct phonological route, bypassing the lexical-semantic stages of processing, the bottleneck only affected the interpreting task.

Because the lexical bottleneck proved to be sufficient to simulate the difference in error rates between shadowing and interpreting, there was no need to add a mechanism to the model to account for syntactic or semantic processing effects. However, the validity of a model for interpreting without syntactic or semantic processing costs may be questioned. In our original study (Chapter 2), we hypothesized that the sentence structure in the narratives used was simple enough to not carry significant syntactic or semantic processing costs. The present study directly tested this hypothesis by removing the narrative context, both syntactic

and semantic, and examining the effect this had on shadowing and interpreting performance. In particular, we removed the function words from our original spoken texts, and then randomised the order of the content words. Participants had to perform interpreting and shadowing tasks on these randomized lists of spoken words. If the deterioration of interpreting performance with increasing speech rate is partly or wholly due to syntactic or semantic processing costs, contrary to what we hypothesized, then we should observe less deterioration of performance with increasing speech rate for the randomized lists of spoken words than we observed previously for the narratives.

3.2 Methods

3.2.1 Participants

Participants were recruited from the Max Planck Institute for Psycholinguistics participant pool. They were native speakers of Dutch, screened using LexTale, an online English vocabulary test, to identify proficient Dutch-English bilinguals. LexTale scores have been demonstrated to be highly correlated with other measures of proficiency (Lemhöfer & Broersma, 2012). Participants with a LexTale score over 85% were invited to participate. In Chapter 2 this was found to be roughly the 70th percentile of the Max Planck Institute participant database, which consists mostly of university students. 20 participants were selected to take part in the study (12 female, mean age 21.4 years). Participants had no prior experience in shadowing or interpreting. Their mean LexTale score was 89.3% (SD = 4.7%), and their self-reported mean age of acquisition was 10.3 years (SD = 1.0 years, N = 16) which approximately coincides with the start of formal English education in most Dutch schools.

3.2.2 Procedure and design

The procedure was kept identical to the procedure used in Chapter 2. Participants were invited to participate in two sessions, one week apart. The first session was a familiarisation session in which participants performed both shadowing and interpreting tasks at various speech rates but data from this session were not analysed. The second session consisted of two blocks of 15 minutes: one shadowing block of five spoken texts presented at different speech rates and one interpreting block of five spoken texts presented at different speech rates. The order of texts, tasks, and speech rates was counterbalanced across participants so that each text was presented to at least two participants at every speech rate and in both tasks, but texts and speech rates were not repeated within participant.

3.2.3 Materials

For the present study, we reused the materials from Chapter 2, breaking up the narrative context by isolating the individual words in the stimulus recordings and then randomising their order. An additional complexity was having to account for function words, which are often hard to process correctly in isolation and function more as syntactic glue for content words. The model in Chapter 2 was found to best account for human error rates when assuming a doubled rate of accumulation for function words (effectively halving the processing time for function words). Since function words comprised approximately half of the words in the original source text, the rate of actual content words per minute was half the speech rate. Semantic information is carried largely in these content words and function words are not easily translated in isolation, therefore we removed the function words, leaving only the content words. The speech rate was then manipulated to keep the rate of content words per minute (cwpm) consistent with the materials used in the previous study (i.e., rates of 50-100 content words per

minute). In addition, repeated words within a text were removed, to ensure there would be no unintentional facilitation from simple word repetition.

3.2.4 Analysis

Participant recordings were transcribed and scored by counting correctly reproduced words. For the interpreting task, a word was scored as correctly reproduced if the translated word was judged to be a correct translation of the target word in the source recording. Since the correct translation of a word taken out of context can be ambiguous, some leniency was used in scoring the translations of these ambiguous words. Additionally, obviously misheard words that were otherwise reproduced correctly were also scored as correct.

A logistic mixed-effects model was used to analyse the effect of speech rate, task, and speech rate-task interaction on participant error rates. The model was constructed using the lme4 package (Bates et al., 2015) and statistical inference computed with the lmerTest package (Kuznetsova et al., 2017) in R (Version 3.3.3, R Core Team, 2013).

3.3 Results and interim discussion

Figure 3.1 shows participants performance on interpreting and shadowing of randomised lists, as well as in the previous study using narratives. As before, both shadowing and interpreting error rates were positively related with speech rate and a speech rate-task interaction effect is visible in the difference in slopes of interpreting and shadowing performance across speech rates. The logistic mixed-effects model demonstrated significant effects of task ($\beta = 0.68, SE = 0.02, p < .001$), source speech rate ($\beta = -0.73, SE = 0.02, p < .001$), and the interaction between task and source speech rate ($\beta = -0.16, SE = 0.02, p < .001$). Random effects of participant and source text were controlled for with random intercepts.

Using a more elaborate random effects structure with random slopes was not possible due to the limited number of observations in each cell.

As visible in Figure 3.1, the slope of interpreting performance is more steeply negative in the randomized materials in the present study than the narratively organized materials in the previous study. The same is true, but less clearly so, for the slope of shadowing performance. Joint statistical analysis of both studies showed that these differences in performance slopes between studies were significant, as evidenced by a content word speech rate by study interaction ($\beta = -0.044, SE = 0.011, p < .001$) in addition to a main effect of study ($\beta = -0.34, SE = 0.011, p < .01$).

Using random lists of words, we replicated the observation of Gerver (1969) and Chapter 2 that performance accuracy decreases more for interpreting than for shadowing with increasing speech rate. If semantic and syntactic structure contributes to the deterioration, then the decrease in performance accuracy with increasing speech rate should be weaker for randomised word lists than for the previously used narratives. However, contrary to this expectation, the decrease in performance was stronger for the random lists than the narratives, which suggests that semantic and syntactic processing costs were entirely absent. That is, we observed that when semantic and syntactic structure is absent, as with the random lists, performance is worse than when semantic and syntactic structure is present. Thus, semantic and syntactic processing yields no cost but helps in interpreting and shadowing.

3.4 Computational model

The observation that interpreting performance is better for narratives than for random lists suggests that the presence of semantic and syntactic structure yields no cost in interpreting. Instead, it seems plausible that the presence of semantic

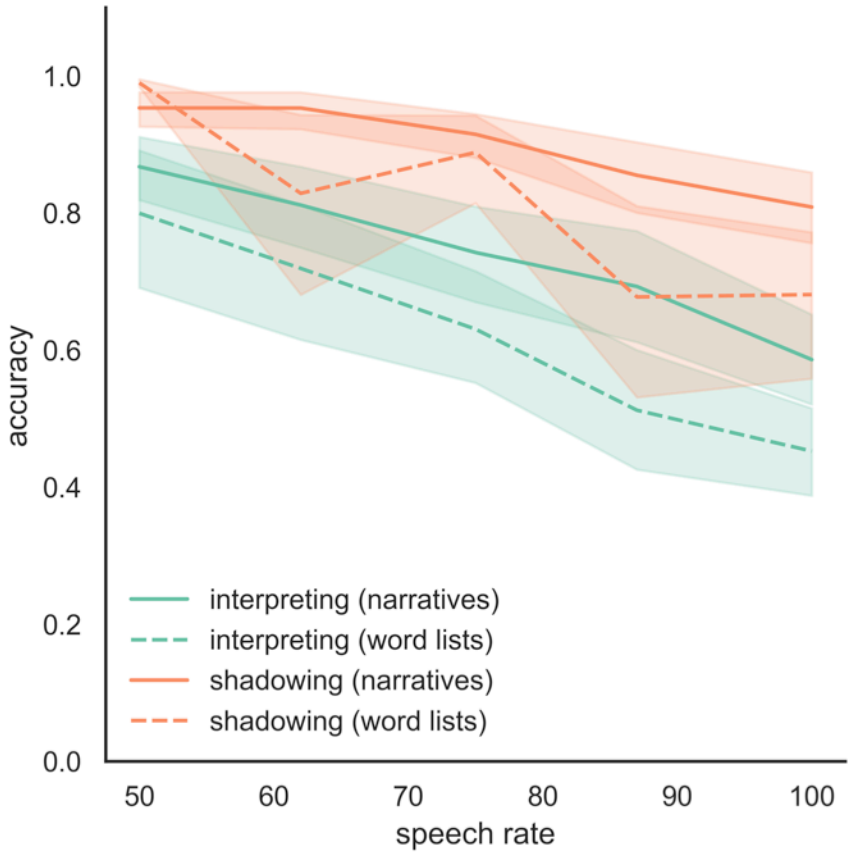


Figure 3.1: Mean proportion of correctly reproduced words while shadowing and interpreting narratives and randomly ordered word lists across source content word speech rates. Narrative shadowing and interpreting performance taken from Chapter 2 for comparison.

and syntactic structure benefits rather than hampers performance. If the presence of semantic and syntactic structure is beneficial for performance, then our lexical bottleneck model may even have underestimated the lexical contribution to the deterioration of the interpreting performance in our previous study. If so, the model should have difficulty fitting the interpreting performance on random lists. We tested this by computer simulations. Following the methodology in Chapter 2, the recordings presented to the participants were processed with Web-MAUS automatic speech segmentation to assign onset times to each word in an orthographic transcription (Kisler et al., 2017). These onset times were then used as input for the computational model, which simulated error rates for shadowing and interpreting.

Running the model with the parameter values used in the previous simulations yielded no decrease in performance accuracy with increasing speech rate for interpreting and shadowing of the lists. For performance to deteriorate, the lexical bottleneck effect should have been much bigger than assumed based on the previous study with narratives. In order to obtain the optimal model fit, the particle swarm optimization procedure (Kennedy & Eberhart, 1995) used in the previous study was repeated with the data collected in the present study. In order to obtain the observed deterioration in performance in interpreting lists (shown in Figure 3.2) the switch cost had to be increased to 520 ms (from 50 ms in the previous study). This suggests that our lexical bottleneck model has underestimated the lexical contribution to the deterioration of the interpreting performance in our previous study. This corroborates our conclusion that the decrease in interpreting performance with increasing speech rate is due to a lexical bottleneck rather than semantic and syntactic processing costs. However, as Figure 3.2 shows, the model predicts shadowing of lists to be entirely unaffected by speech rate, whereas the empirical data show that shadowing also deteriorates. Increasing the lexical switch cost does not affect shadowing because most of the process-

ing in our simulation of the shadowing task flows though the direct phonological pathway (the shortcut route). This suggests that this direct route is also subject to a bottleneck.

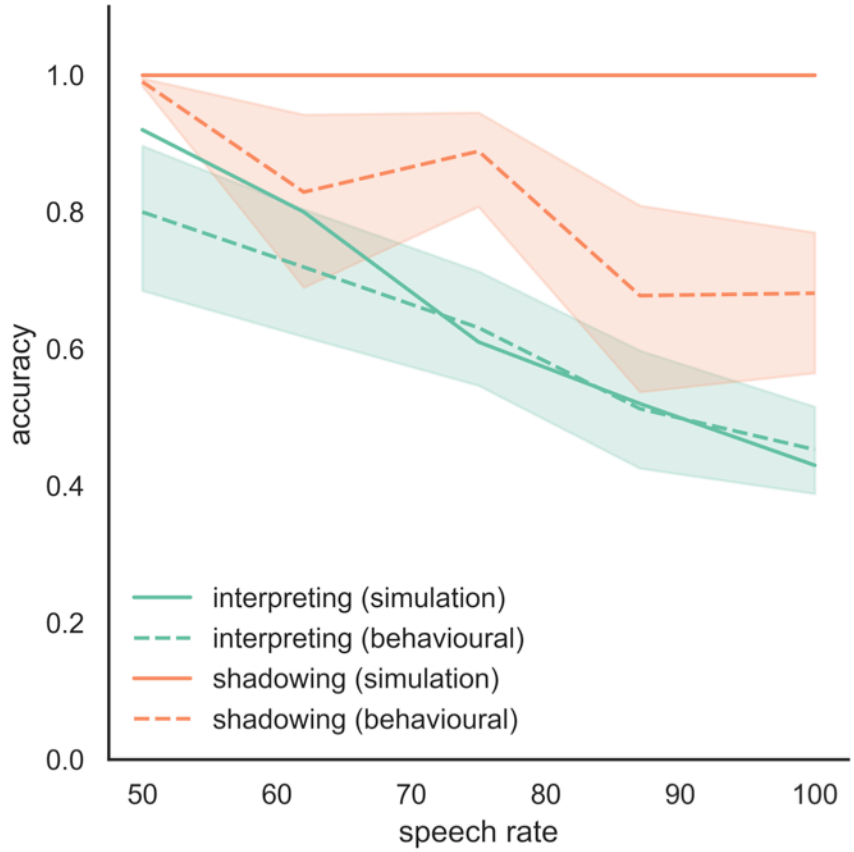


Figure 3.2: Mean proportion of correctly reproduced words while shadowing and interpreting across source speech rates in the behavioural experiment and model. Shaded areas represent standard error for the behavioural data.

3.5 Discussion

While the interaction between speech rate and task observed by Chapter 2 is replicated, there are key differences between the results reported in that study and the results reported above. In particular, the slope of interpreting performance is more steeply negative in the randomized materials in the present study than the

narratives in the previous study. The same is true, albeit less clearly, for the slope of shadowing performance. This indicates that semantic and syntactic processing costs do not contribute to the deterioration of performance with increasing speech rate, which supports our hypothesis of a lexical bottleneck.

However, the direct comparison between the present and previous study is somewhat complicated by the distinction between speech rate in words per minute and content words per minute. As discussed above, in constructing the materials it was decided to not include function words because they are not trivial to translate out of context. This left only content words, functionally doubling the density of semantic information in the materials. To compensate, speech rate was adjusted by taking the speech rate in content words per minute from the previous study, which comes out to roughly half the speech rate in words per minute. In terms of semantic information density this may be roughly comparable, but it is plausible that pre-lexical speech comprehension processes are more sensitive to the raw speech rate than the semantic information density-corrected speech rate. If anything, halving the speech rate from the previous study to arrive at content words per minute likely underestimates the difficulty of those materials compared to the materials used in the present study. Therefore, the difference in slopes between the previous study and the present one that is visible in Figure 3.1 likely represents a lower bound for the actual difference. Thus, our conclusion that the decrease in performance was stronger for the random lists than the narratives is warranted.

Our empirical findings and computer simulations suggest that the difference in slopes between studies represents a facilitation effect from narrative context, as opposed to any cost due to semantic or syntactic processing. Our assumption in Chapter 2 that semantic and syntactic structure building is not a significant bottleneck in these simple narratives therefore seems justified. If the presence of semantic and syntactic structure has a facilitatory effect on performance, then

our lexical bottleneck model may even have underestimated the lexical contribution to the deterioration of the interpreting performance in our previous study.

There is evidence for such facilitatory feedback into early speech processing from both syntactic and semantic context information (Tyler et al., 2002; Zwitserlood, 1989) and differing amounts of contextual facilitation in the form of feedback would not only explain the difference in interpreting performance between narratives and word lists, but could also help explain the difference in shadowing performance between the previous and the present study. In the previous study, shadowing performance was largely preserved at higher speech rates. We posited that shadowing occurred via a faster, low-level "shortcut" route, bypassing the lexical bottleneck. This route is possible for shadowing because when literally repeating a word, an output phoneme can be selected immediately after recognizing the input phoneme (Roelofs, 2004, 2014). When shadowing using this route, conceptual integration occurs only after production of each lexical item. Based on this assumption, our computational model predicted extremely low error rates for shadowing in the present study, which proved to be erroneous. This may be partly due to comparatively less use of the purely phonological route and more use of the full conceptual-semantic route in the present study. However, even if full conceptual integration in shadowing is post-hoc, early conceptual-semantic information could feed back into lower level comprehension processes during both shadowing and interpreting tasks. Such feedback would provide a form of continuous semantic priming. Additionally, via a similar mechanism, syntactic effects could feed back from the lemma level into lower level comprehension processes, priming words based on grammatical features in a narrative context, but not when shadowing or interpreting word lists. Marslen-Wilson (1985) found evidence that such context effects are not only present in shadowing, but that they are stronger at the relatively long latencies we observed in the present study (i.e., the effect is stronger when there is more time for conceptual-

semantic and syntactic information to feed back down to lower-level processes). Semantic and syntactic context effects were also observed in a phoneme restoration task, but only if the first syllable of the affected word was left intact, indicating a direct effect of sentential context on phonological processing.

Assuming a small switch cost, these feedback mechanisms may perhaps further explain why the computational model from Chapter 2 does not, in its current form, account for decreased performance in the absence of contextual (syntactic and/or semantic) facilitation. The model operates in a feedforward manner, allowing it to simulate contextual facilitation of conceptual integration, which would allow it to partially model contextual facilitation of interpreting. However, its feedforward architecture does not allow it to simulate the conceptual-semantic information feeding back and facilitating lower level processes in a manner that would plausibly cause contextual facilitation during shadowing.

A model that is capable of this facilitation would entail a substantially higher number of parameters, due to the need to tune the feedback connections and contextual facilitation. Fitting such a model to the current dataset of behavioural error rates is likely to result in an overfitted model; therefore we do not attempt it here. However, if additional behavioural data were used to account for the contextual facilitation parameters, the number of parameters that would need to be fitted in an expanded model could be reduced. This would address the overfitting problem, but such a model would likely still need an additional bottleneck to account for the difference in performance decline with increasing speech rate between shadowing and interpreting observed in the behavioural data. Based on the available evidence from other experiments, (e.g., Cook & Meyer, 2008; Piai et al., 2014) the lexical response selection bottleneck we suggested in Chapter 2 remains a likely source of this speech-rate task interaction.

3.5.1 Conclusion

In Chapter 2, we proposed that a lexical bottleneck is sufficient to explain the difference in speech rate-induced errors between shadowing and interpreting tasks. In the present study, we directly tested our lexical bottleneck hypothesis by examining interpreting and shadowing performance on randomised word lists. We replicated the stronger decrease of interpreting than shadowing performance with increasing speech rate. The decrease for the lists was even stronger than previously observed for narratives, which suggests that semantic and syntactic structure costs are absent. This suggestion was partially supported by computer simulations, which suggest that our lexical bottleneck model even underestimated the lexical contribution to the deterioration of the interpreting performance in our previous study. However, the observed deterioration in shadowing accuracy cannot be explained using only a lexical bottleneck. Contextual facilitation effects would explain the discrepancy, but processes producing these effects could not be modelled in the present study. We conclude that a lexical bottleneck remains plausible, and that to fully explain the deterioration of interpreting and shadowing performance with increasing speech rate in both narratives and randomized word lists, semantic and syntactic context must facilitate processing, rather than inducing a processing cost.

4 | Lexical and contextual factors facilitate concurrent speech comprehension and production in simultaneous interpreting and shadowing[1]

Abstract

The time course of speech processing in simultaneous interpreting and shadowing of narratives points to a lexical bottleneck, as shown in computer simulations in Chapter 2. Our simulation model accounts for the competing needs of comprehension and production for lexical processing by assuming a bottleneck switching mechanism with an associated switch cost. However, the model was found to not generalize well from stories to randomized word lists (Chapter 3). We hypothesized that the model failed to generalize because it did not include facilitation from narrative context. Here, we investigate the nature and magnitude of contextual facilitation effects for narratives as opposed to randomized word lists in speech shadowing and simultaneous interpreting. We analyzed the speech latencies using Bayesian regression with sparsity-inducing priors to select relevant priors from a variety of lexical and contextual factors including phonological length, cognateness, frequency effects, transitional probabilities at the word and lemma level, syntactic complexity, and similarity to semantic context. The regression coefficients reveal differences between interpreting and shadowing that indicate differences in planning scope and semantic processing. Comparison of latencies in narratives and randomized word lists indicates that contextual cues strongly facilitate concurrent speech comprehension and production.

[1] Adapted from Van Paridon, J., Alday, P. M., Roelofs, A., & Meyer, A. S. (in prep.). Lexical and contextual factors facilitate concurrent speech comprehension and production in simultaneous interpreting and shadowing.

4.1 Introduction

Errors in simultaneous interpreting and shadowing of spoken narratives vary as a function of speech rate (Gerver, 1969). In computer simulations, we demonstrated that the varying accuracy of interpreting and shadowing performance can be explained by assuming a lexical bottleneck (Chapter 2). Simultaneous interpreting requires lexical processing for both comprehension and production, which we accounted for in the model by switching lexical processing between comprehension and production, with a small switch cost incurred every time a switch is made. This model was able to closely fit error rates in the interpreting and shadowing of narrative text, using a processing hierarchy and parameters based on a review of the psycholinguistic literature (Indefrey & Levelt, 2004). In Chapter 3, however, we found that the model did not generalize to error rates in interpreting and shadowing randomized word lists. We hypothesized that the model was unable to correctly fit to error rates on randomized word lists because these lists do not provide the facilitatory context (be it semantic, syntactic, or both) that the narratives used in Chapter 2 provide. Since our mechanistic model operates only on the timing of the spoken input and not on linguistic content of said input, it did not explicitly account for such contextual facilitation effects in Chapter 2, and therefore the fitted parameter values do not generalize to the data collected in Chapter 3, where the contextual facilitation is absent.

Rather than revisiting the mechanistic model and implementing an ad hoc contextual facilitation fix, we focus in the present study on examining the locus and magnitude of the hypothesized contextual facilitation effect. By contrasting semantic and syntactic context effects, we aim to uncover at which processing stages contextual facilitation occurs during interpreting and shadowing, and how this facilitation affects the difficulty of each task. This information will allow us to make a principled adaptation to future mechanistic modeling, accounting for

contextual facilitation and potentially improving the fit to error rates in the data collected in Chapter 3.

4.1.1 Modeling of speech shadowing and simultaneous interpreting

Earlier models of interpreting have mostly been process models (see e.g., Moser, 1978) or models relating various component skills to interpreting (see e.g., Christoffels et al., 2003). But while these models are useful in understanding how concurrent comprehension and production of speech are coordinated during simultaneous interpreting, they largely do not make specific quantitative predictions about language processing and task performance in simultaneous interpreting and speech shadowing. A notable exception is the work of Seeber and Kerzel (2012) relating differences in word order between source and target language to pupillometry data (as a proxy for cognitive effort or working memory load) but the predictions they tested were fairly narrow, pertaining only to word order differences between closely related languages.

In Chapter 2 we proposed a computational model of speech shadowing and simultaneous interpreting error rates based on speech comprehension and production processing times originally compiled in a review of the word comprehension and production literature by Indefrey and Levelt (2004). While the proposed model could fit the observed error rates well, it has several limitations. Most importantly, it only predicts error rates at the group mean level and it is blind to the linguistic (semantic and structural) content of the speech. This makes judging the theoretical validity of the process model difficult. The processes seem to be modeled correctly in terms of order and relative duration, but they are based on studies of single word and noun-phrase production. Are we therefore overlooking effects of processing linguistic content, particularly at the scope of sentences, rather than single words?

In contrast to Chapters 2 and 3, in which we use a mechanistically inspired model, in the present study we aim to descriptively model the same tasks at the level of input-output speech latency, in order to elucidate the linguistic factors, both at the purely lexical level and at the contextual level, that affect speech shadowing and simultaneous interpreting performance. We expect lexical facilitation effects to occur in both the narrative and randomized word list conditions, while contextual effects should only occur in the narrative condition. Furthermore, the different processing hierarchies we propose for shadowing and interpreting in Chapter 2 lead us to expect that semantic and syntactic effects in shadowing, if they occur at all, should be smaller than in interpreting. This is because semantic and syntactic processing are not critical to performing the shadowing task (but potentially do occur post production of a word and could therefore still have some effect on the recognition and production of subsequent words). By uncovering the locus of contextual facilitation effects, we can better understand where our computational model falls short, and perhaps amend it.

4.1.2 Factors potentially affecting interpreting and shadowing latencies

It seems clear that many factors affect speech latencies in simultaneous interpreting and speech shadowing, but tightly controlled experimental paradigms generally afford testing only one or two factors at a time. A rich, naturalistic dataset on the other hand allows for the simultaneous modeling of various facilitatory and delaying effects at both the lexical level and the contextual level (see e.g., Alday, 2019; Alday et al., 2017, for application of this concept in electrophysiology of language). In the present study, we examine the effects of cognateness (orthographic and phonological), word length (syllabic and phonemic), word frequency, word-level transitional probabilities (forward and backward, for bigrams and trigrams), transitional probabilities from lemma to part-of-speech

(grammatical category), similarity to semantic context (across different context windows of varying lengths), and syntactic dependencies (left and right children). Next, we briefly describe these variables and our motivation for including them.

Cognateness

Dutch and English are closely related Germanic languages and consequently there are many cognates: Many English words have the same etymology as their Dutch translations, resulting in significant similarity in both phonology and orthography. Cognateness has been shown to facilitate production in bilinguals (Costa et al., 2000) and ERP studies of the time course of cognate production trace this effect to the phonological processing stage (Christoffels et al., 2007). However, Strijkers et al. (2010) also observe cognateness effects at an earlier stage, in lexical access, which they attribute to co-activation of lexical representations in both languages (effectively raising the frequency of activation for a lexical item above the frequency observed in a single language). They hypothesize this co-activation is mediated by indirect links between lexical items through shared phonological segments (see Costa et al., 2005, for review). The elevated effective word frequency hypothesis is speculative, but regardless of whether one accepts it, it is plausible that any facilitatory effect of cognateness must arise through (indirect) links at the level of phonological representations. This suggests there are two possible mechanisms by which cognateness is likely to facilitate simultaneous interpreting in the present study: Through facilitated word recognition and production (perhaps because of the raised effective frequency) and through indirect links between phonological representations, bypassing the need for conceptual-semantic mediation in translating.

Quantifying cognateness, however, is not straightforward. The Levenshtein (edit) distance between phonetic transcriptions (in IPA) of the phonological words is often larger than the perceived difference between words, because

of systematic (and therefore largely transparent) shifts in phoneme-grapheme correspondences between English and Dutch that occurred as these languages diverged over time. Orthographic Levenshtein distance can therefore paradoxically be a better (although still noisy) measure of perceived cognateness. Levenshtein distances are correlated with word length, however we account for this by normalizing the Levenshtein distance for word length (or, to be more specific, the mean of the lengths of both the English and Dutch words). We include both phonetic and orthographic length-normalized Levenshtein distance as predictors in the present study.

Word length

Length is a lexical factor we might expect to have a delaying effect, for two reasons: In comprehension, the uniqueness point of longer words is generally later than that of short words, meaning that recognition of long words is likely to take longer than recognition of short words. Surprisingly, evidence that the uniqueness point of a word affects speech latency in a continuous speech task is mixed, with two studies on speech shadowing finding that uniqueness point affects speech latency only at slow presentation rates, if at all (Radeau & Morais, 1990; Radeau et al., 2000). In production, uttering a longer word requires more planning and preparation, but whether this affects onset latency depends on whether planning and preparation occur wholly before the onset of production. Evidence for effects of word length in speech production is mixed, and whether length effects are observed seems, in practice, to be dependent on task demands such as response deadlines (Meyer et al., 2003).

We can measure word length in number of phonemes and number of syllables. There is some evidence that articulatory preparation occurs at the syllable level, making number of syllables particularly relevant as an index of planning difficulty, but again, this is only relevant if participants in the present study plan

multiple syllables before they start articulating (Levelt & Wheeldon, 1994; Meyer, 1990). Unfortunately, since most phonemes map onto a single grapheme (or at most two, with rare exceptions), and syllables tend to have a somewhat consistent number of phonemes, these two measures are highly collinear and their effects are therefore impossible to disentangle in our data.

The aforementioned similarity between English and Dutch has implications for disentangling lexical comprehension and production effects as well: the high number of cognates means that for many words, length and frequency in English tends to be very similar to length and frequency in Dutch. This leads to a somewhat vexing situation where, although simultaneous interpreting theoretically means that perceived lexical items are different from produced lexical items, the specifics of our selected language pair (selected for the relatively good availability of proficient bilinguals) make it difficult to distinguish between a comprehension or production locus for frequency and length effects.

Word frequency, ngram frequency, and transitional probability

That word frequency facilitates speech perception has been generally accepted for well over half a century (see e.g., Rosenzweig & Postman, 1958). In speech production, word frequency has a similar facilitatory effect, generally thought to occur at the level of phonological encoding (Jescheniak et al., 2003; Meyer et al., 1998). We therefore expect words that occur more frequently to be both recognized and produced more easily. More recently, such facilitatory effects have also been observed at the phrasal level in both production (Shao et al., 2019) and comprehension (Arnon & Snider, 2010). Further complicating these findings, there is some evidence from eye movements to suggest that what readers are truly sensitive to is not frequency, but rather transitional probability (McDonald & Shillcock, 2003). Transitional probability in this context means the probability that a word will occur given the preceding context (forward transitional probability, or FTP)

or that a certain context has preceded a given word (backward transitional probability, or BTP). Like frequencies, these quantities are log-transformed for use as predictors (note that in the reading and natural language processing literature, it is common to take the negative logarithm of transitional probability, either predicted or observed, and call it surprisal, see e.g., Frank, 2013).

As discussed in Chapter 5, the common practice of taking the logarithm of the transitional probabilities results in a notable equivalence. Dividing the bigram probability by the word probability (or 1-back word probability) is equivalent to a subtraction on the log scale. This essentially means that if we include bigram frequency, BTP, and word probability as predictors, we create a multicollinearity problem. Including the full set of frequency and probability measures is simply not possible, because it renders the model unidentifiable. Note that these same concerns apply to trigrams.

In order to obtain estimates of word, bigram, and trigram frequencies in conversational English and Dutch, we used the frequencies compiled in Chapter 6, tabulated from OpenSubtitles, a large archive of (pseudo-conversational) transcribed film and television subtitles. Using these word, bigram, and trigram frequencies, we can compute relevant conditional probabilities. In addition to transitional probabilities for word n-grams, we also compiled lemma to part-of-speech transitional probabilities: the probability that a lemma or lemma bigram is followed by a particular class of word. These probabilities can be understood to reflect potential pre-activation of a word class (through lemmas) facilitating lexical selection.

Because of the aforementioned perfect collinearity issue described, we only include the logarithm of bigram and trigram FTP and BTP (i.e., bigram and trigram surprisal). These have the most natural interpretation in the context of speech tasks: If I have previously observed these words, which word comes now? And, if I am observing this word now, which words are likely to have preceded it? Con-

versely, using whole bigram and trigram frequencies ignores the compositional nature of language.

Semantic context

Semantic similarity of a word to the context preceding it can facilitate comprehension and production. This can be observed in single-word semantic priming studies, but in the present study a longer preceding context can be modeled, allowing us to juxtapose the relative effects of distal and proximal semantic context. We operationalize semantic context using subs2vec (Chapter 6), a word2vec-style co-occurrence model implemented using the fastText algorithm (Bojanowski et al., 2017). Word vectors produced by this co-occurrence model can be averaged to construct semantic context models of any phrase length. Simple semantic context models of this type have previously been used in EEG studies to show that similarity to preceding semantic context facilitates word recognition in a top-down fashion (Broderick et al., 2018; Broderick et al., 2019). Such top-down facilitation, if we can also observe it behaviorally (i.e., in our dataset of speech latencies) is in line with our hypothesis that narrative text processing is subject to contextual facilitation effects, while processing of word lists is not (Chapter 3). The combination of semantic context models and EEG has also been used to differentiate semantic context effects from simple word predictability, demonstrating that these measures produce distinct facilitatory effects (Frank & Willems, 2017). More generally, the semantic information encoded in simple semantic context similarity models has been found to be sufficiently predictive of brain activity patterns generated by comprehension of narrative text that perceived sentences can be decoded from fMRI data with some success (Pereira et al., 2018).

In the present study, we compute cosine similarity between the word vector for the current word and the context vector, a measure which does not explicitly represent priming, spreading (pre-)activation, prediction, or ease of conceptual

integration, but rather the amount of semantic information from the preceding context that could potentially cause facilitation or delay through various different mechanisms.

Cosine distance to preceding context is related to, but distinct from ngram frequencies and transitional probabilities in that it is a more specific (though by no means perfect) measure of semantics. Frequencies and transitional probabilities encode both phrasal structure and semantics to a certain degree. For instance, the will probably be followed by a noun, and drive to will often be followed by a city or other type of place. Backward transitional probability, in particular, has been proposed as an index of conceptual integration, but cosine distance actually quantifies the conceptual gap that has to be closed to integrate a word into the preceding context, without encoding the specific word order in the way that transitional probabilities do.

Syntactic complexity and predictability

Words have syntactic relationships with the words surrounding them, in some cases having dependencies to their left or right that need to be taken into account to interpret a given word correctly. The requirement to keep track of these dependencies in order to comprehend a sentence and produce a correct translation means that some sort of representation needs to be held in working memory. The increased working memory load associated with a tracking a high number of right-branching dependencies could cause delays in planning for production. Additionally, in simultaneous interpreting, differences in word order between the source and target language have been shown to increase cognitive load, as measured by pupil dilation (Seeber & Kerzel, 2012). For the simple sentences used in the present study, however, both the source (English) and target (Dutch) languages predominantly have the same (subject-verb-object) word order, which should minimize the effect that word order has on speech latencies. We parsed

the source narratives using the dependency parser implemented in SpaCy (Honnibal & Johnson, 2015) to obtain the number of left- and right-branching dependencies for every word for use as predictors in our model (but for an alternative approach to quantifying dependencies in interpreting, see Liang et al., 2017).

4.2 Method

This study concerns input-output speech latencies and accuracy annotated on participant recordings originally collected in Chapters 2 and 3, which both modeled error rates, rather than speech latencies. The data collection procedure used in these prior studies will be reproduced here for completeness.

4.2.1 Participants

Forty-one highly proficient Dutch-English bilinguals (native speakers of Dutch) were recruited from the Max Planck Institute for Psycholinguistics participant database. These participants represent the 70th percentile of an initial cohort of 215 which was screened for English proficiency using the LexTale vocabulary text (Lemhöfer & Broersma, 2012). For more details on participants see Chapters 2 and 3.

4.2.2 Procedure and design

Participants were asked to shadow five texts and interpret five texts, with short breaks in between. Order of shadowing and interpreting blocks was counterbalanced across participants. Half the participants were presented with narrative texts, the other half with randomized word lists. The speech rate at which texts were presented was varied in five steps from 50 to 100 content words per minute. In the narrative condition, this results in a real speech rate of 100 to 200 words per minute. In the randomized word list condition function words were removed, re-

sulting in the real speech rate matching the content word speech rate of 50 to 100 words per minute. The order of presented speech rates was counterbalanced across texts and participants.

4.2.3 Materials

Both the narrative and the randomized word lists were based on samples from books targeted at children aged six to ten years old. We used ten samples of approximately 300 words as the narrative texts, and randomized the word order after removing function words to create vocabulary-matched randomized word lists from those same samples. Narrative samples were recorded at a fixed speech rate by a native speaker of English using a custom teleprompter script, randomized samples were cut from these recordings. We collected and, where necessary, computed relevant predictors for the materials. Correlations between these predictors are reported in Figures 4.1 and 4.2.

4.2.4 Annotating participant recordings

Participant recordings were first transcribed to get an orthographic representation of participants utterances. We then forced-aligned the participant recordings to the transcripts using WebMAUS (Kisler et al., 2017) to find the onset and offset of each individual word in the transcripts. With the English source text and participant utterances aligned, lapses and incorrect responses were rejected word-by-word. This annotation procedure produced more than 50,000 input-output speech latencies to be used as a dependent variable in our analyses.

4.2.5 Modeling input-output latencies

To investigate the effect on the input-output latency in the interpreting and shadowing tasks of the various predictors we have constructed, we construct regres-

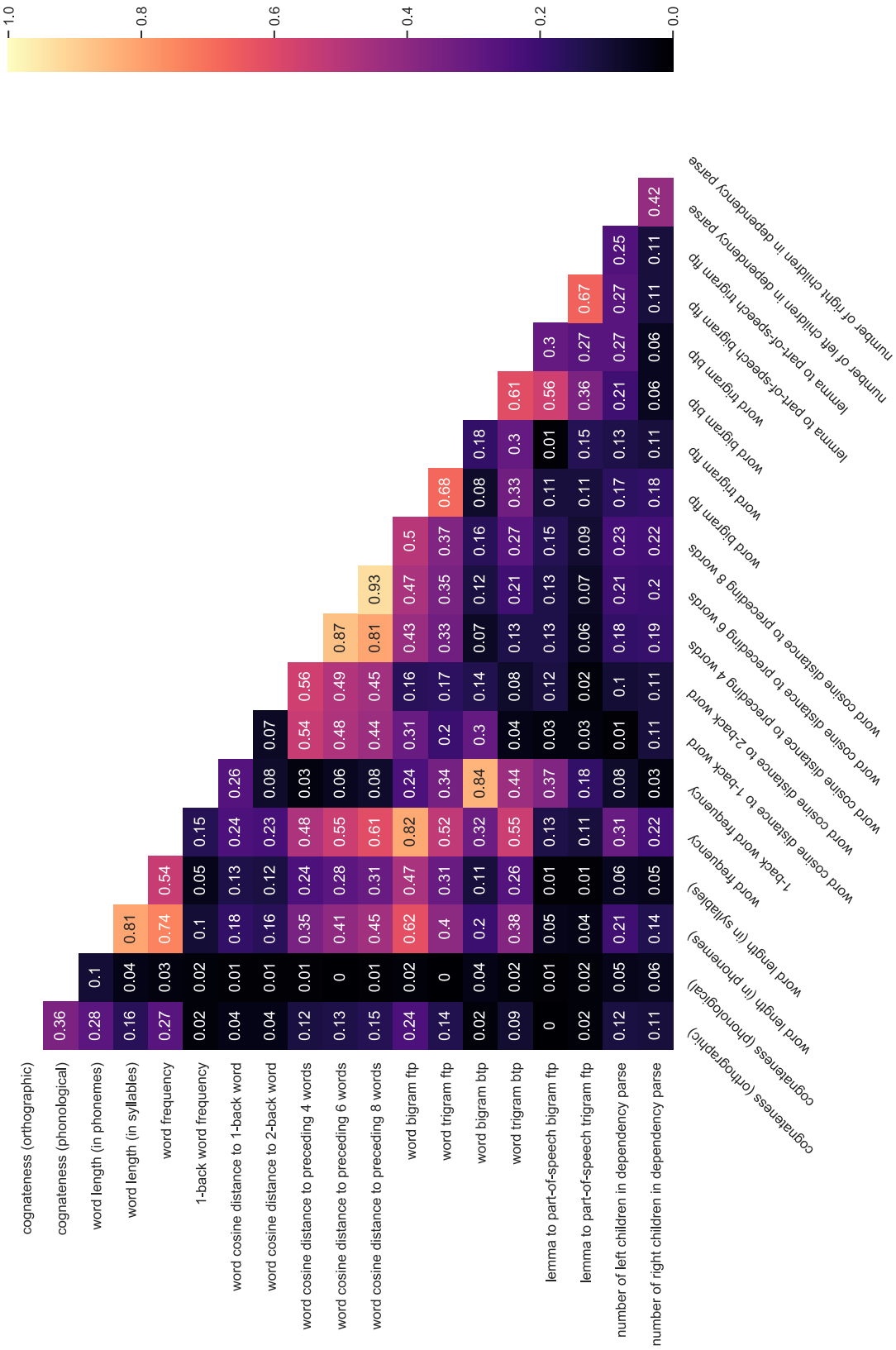


Figure 4.1: Heatmap of absolute correlations between predictors in the narrative condition.

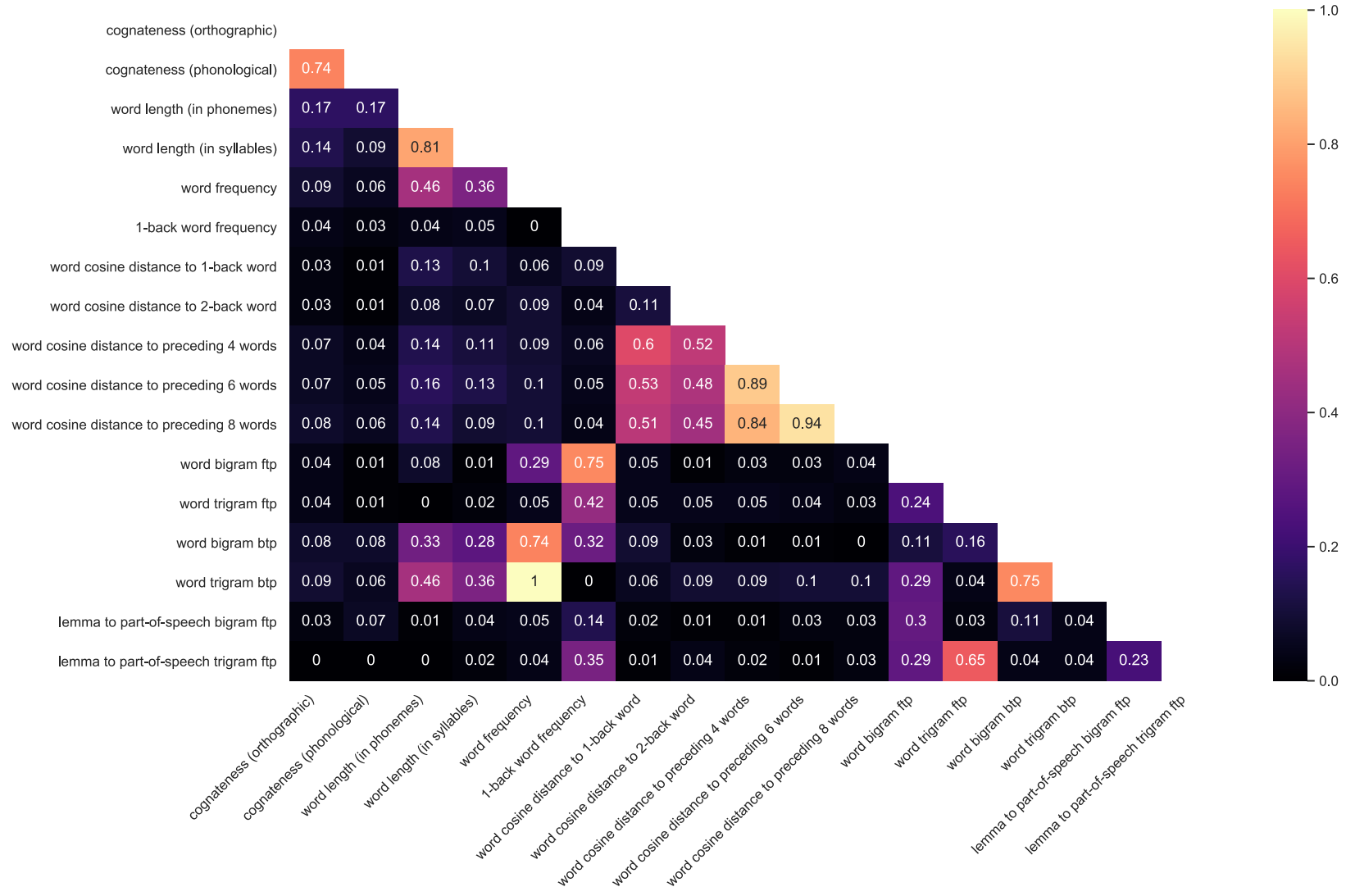


Figure 4.2: Heatmap of absolute correlations between predictors in the randomized word list condition.

sion models. In constructing these models, we have to make a number of assumptions and decisions, of which some, but not all, are generally accepted practice in psycholinguistics. Most consequential are our choice of parameter selection method (sparsifying priors) and dependent variable transformation (log-transform), which we will therefore briefly motivate.

Encouraging coefficient sparsity

We have a large number of predictors, some of which are exploratory (i.e., we do not have a good a priori hypothesis for the magnitude of their effect) and many of which are collinear to some degree (meaning we could not attribute variance in latencies uniquely to one predictor in a regression). We want to encourage sparsity, shrinking the coefficients (and consequently, our effect size estimates) for irrelevant predictors to zero. In a Bayesian regression framework, sparsity can be encouraged by selecting appropriate priors. The most convenient sparsity prior is the Laplace prior, which is functionally equivalent to the frequentist practice of setting an L1-penalty on the total magnitude of the regression coefficients, better known as Least Absolute Shrinkage and Selection Operator (LASSO, see Tibshirani, 1996). However, Laplace priors only encourage sparsity by having relatively more of their probability mass close to zero than Normal priors (equivalent to ridge regression/L2-penalty, or mild shrinkage) do; because the Monte Carlo methods used for Bayesian statistics are only exact in their asymptotic behavior and the shrinkage is smooth, they generally do not yield an exact numerical zero for the posterior mean.. In practice, Laplace priors behave more like extra strong shrinkage priors: over-regularizing relevant coefficients while not shrinking irrelevant coefficients to exactly zero, neither of which are desirable properties. Various improvements over the Laplace prior have been suggested, of which the Finnish Horseshoe (Piironen & Vehtari, 2017c) most exhibits the behavior we seek. The original Horseshoe prior (Carvalho et al., 2010) has the

unusual property of having infinite density at zero, but still having flat tails, resulting in coefficients either shrinking to zero, or exhibiting no shrinkage at all. Some shrinkage of non-zero parameters is actually desirable in case of a weakly identifiable likelihood, so that non-zero parameters do not take on arbitrarily large values (Piironen & Vehtari, 2017c). The improved Finnish Horseshoe can be conceptualized as a combination of several priors: a Horseshoe prior on each parameter encouraging small estimates to shrink to zero, a mild shrinkage prior (roughly equivalent to a Student T prior) encouraging shrinkage of non-zero parameters, and a global prior reflecting our expectation for the overall number of non-zero parameters (Piironen & Vehtari, 2017b).

As a more general point, in high-dimensional models such as the ones we are reporting here, selecting parameters through model selection (whether through Bayes Factors or indices like WAIC and PSIS-LOOIC) is computationally expensive (if not intractable) while selecting parameters through sparsity has the advantage of being computationally tractable and, if using Finnish Horseshoe priors, functionally equivalent to performing Bayesian model averaging over all possible nested models (Piironen & Vehtari, 2017a).

Log-linear input-output latency

As visible in Figures 3 and 4, our dependent variable, speech latency, is highly skewed. While the modal speech latency is less than two seconds in every task, there is a long tail of higher speech latencies. We take the logarithm of speech latency as dependent variable in our modeling, shown in Figure 5, for two reasons, one statistical and one conceptual. From a statistical point of view, linear models carry the implicit assumption that model error (i.e., residuals) is normally distributed. Log-transforming a skewed normal distribution so that it becomes normal generally improves the normality of the model error as well. Furthermore,

log-transforming the input-output latency carries the implicit assumption that effects are multiplicative, rather than additive:

$$\log(a * b) = \log(a) + \log(b) \tag{4.1}$$

Conceptually, the multiplicative assumption makes more sense than the additive assumption: In the additive case, if multiple strongly facilitative factors apply to a particular word, a model could potentially predict an unreasonably small or even a negative input-output latency. In the multiplicative case, however, effects are relative. This fits with the intuition that a small change in a short latency is more consequential than an equally small change in a long latency. We are effectively saying that for every single standard deviation increase in word length, we might expect a 5% increase in input-output speech latency, as opposed to a 100 millisecond increase in latency. Consequently, even for a word to which multiple facilitating factors apply, this cannot result in a negative predicted latency. Raw latencies are reported in Figures 4.3 and 4.4, and standardized log-latencies in Figure 4.5.

Regession model sampling

The Bayesian multilevel linear regression models were implemented in PyMC3, the probabilistic programming package for Python (Salvatier et al., 2016). Models were estimated by Markov Chain Monte Carlo (MCMC) sampling, using the No-U-Turn Sampler (NUTS) (NUTS, see Hoffman & Gelman, 2014). Markov Chain starting points were obtained with Automatic Differentiation Variational Inference (ADVI, see Kucukelbir et al., 2017). Ten chains were run for 4000 tuning samples, followed by 1000 samples per chain for a total of 10,000 posterior samples.

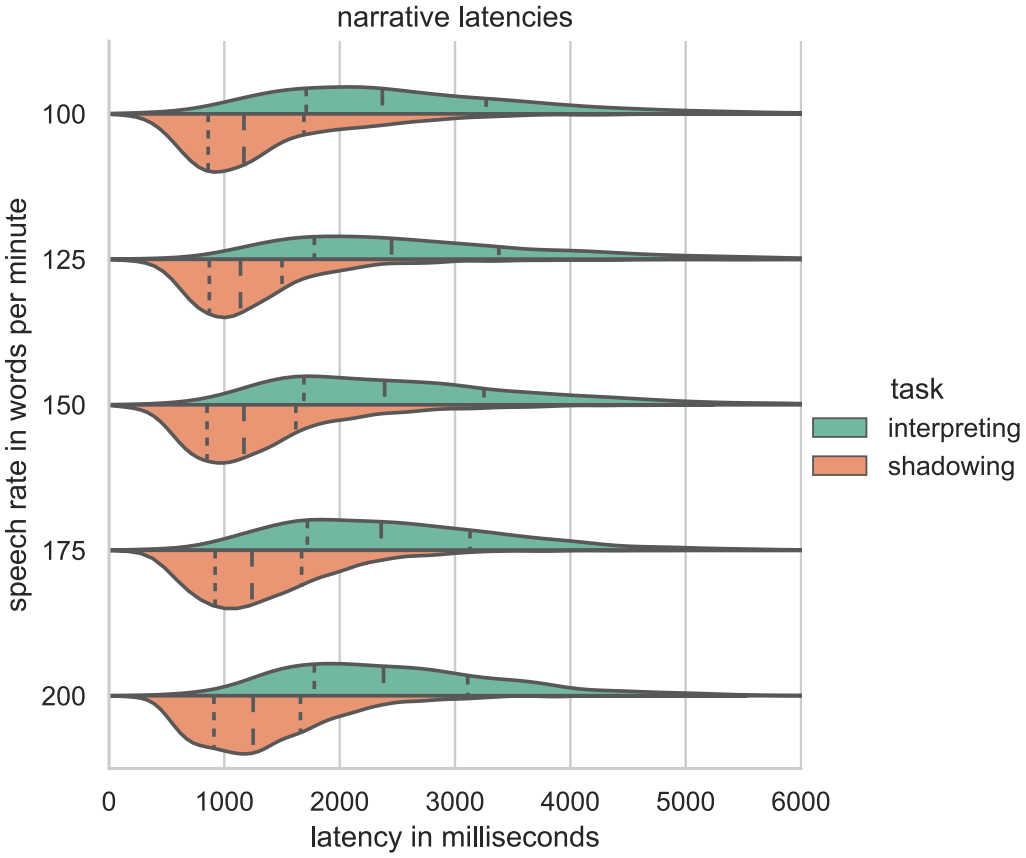


Figure 4.3: Raw speech latencies in milliseconds when shadowing or interpreting narratives.

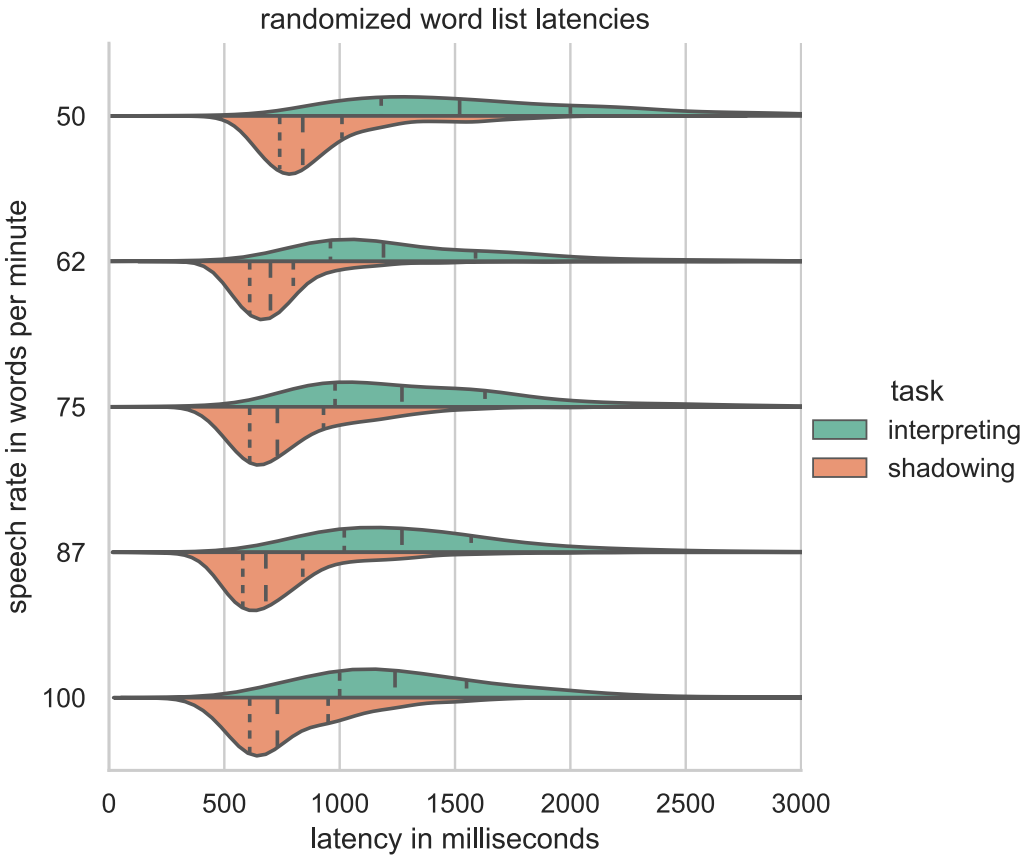


Figure 4.4: Raw speech latencies in milliseconds when shadowing or interpreting randomized word lists.

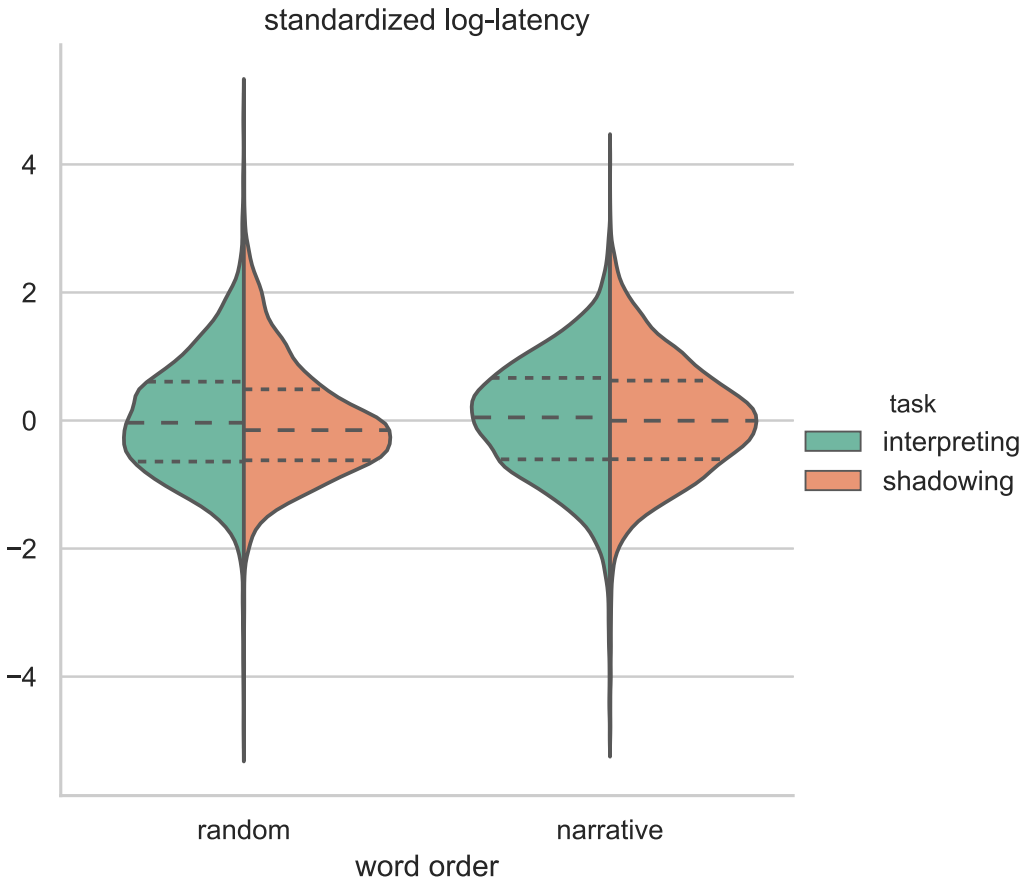


Figure 4.5: Standardized log-latency of speech latencies, dependent variable for the analyses reported here.

4.3 Results

4.3.1 Response latencies

Before examining the results of the regression analysis, it is worth exploring what the basic distribution of response latencies in each condition reveals about what size chunks participants are holding in working memory while performing the task. In the narrative condition (see Figure 4.3), the median response latency when interpreting is around 2400 ms, regardless of speech rate. This corresponds with 4 words at the low end of speech rate (100 wpm) and 8 words at the high end. When shadowing, however, the median response latency is around 1200 ms, corresponding with 2 words at the slowest speech rate, and 4 words at the highest speech rate.

In the randomized word list condition (see Figure 4.4), median interpreting latency is around 1200 ms, except at the lowest speech rate, 50 wpm, where it is roughly 1500 ms. This corresponds to 1.3 words at the lowest speech rate, and 2 words at the highest speech rate. Median shadowing latency in this condition is around 700 ms, except at the lowest speech rate, where it is roughly 850 ms. This corresponds with 0.7 words at the lowest speech rate, and 1.2 words at the highest speech rate.

While there are clear differences in latency between conditions, there is notably little variation in latency distribution across speech rates within each condition. This seems generally inconsistent with a strategy of capturing a given amount of words before shadowing or interpreting them.

Regression model sampling diagnostics

Figure 4.6 shows the coefficient estimates and their credible intervals from sparse Bayesian regression. Estimating a single, pooled model for all four conditions

proved difficult. The combination of difficult posterior geometry due to the sparsifying priors and a large number of predictors plus interactions and second order interaction makes a pooled model hard to sample from. In the coefficient plots the four condition are contrasted by color, but note that the narrative data and randomized word list data come from separate groups of participants, while the interpreting versus shadowing contrasts are essentially within-participants. The sparsity priors combined with the contrast coding used for the narrative versus word list and shadowing versus interpreting conditions result in a model that effectively assumes that most effects will be zero, but also that most effects will be the same across conditions. Where effects differ between conditions, it functionally means there was additional evidence to support a first- or second-order interaction effect over and above the main effect of the predictor.

There were some divergences during sampling, on average 20 per chain for each model. Finnish Horseshoe priors, while having desirable properties with respect to sparsity and shrinkage, are notably difficult to sample from. Betancourt (2018) notes that better exploration of the funnel can be achieved by tuning NUTS parameters (increasing acceptance probability and maximum tree depth). After increasing the acceptance probability from 0.8 to 0.99, and the maximum tree depth from 10 to 15, the number of divergences decreased by an order of magnitude. This, in combination with visual exploration of posterior pair plots, leads us to conclude that divergences during sampling are due to difficulty exploring the high-dimensional funnel around posterior modes which shrink to zero, and not some other, more problematic pathology. This makes it unlikely that our inference is biased to any meaningful extent by these divergences.

4.3.2 Differences between conditions

Uncertainty in estimates is generally somewhat larger for the randomized word list condition than for the narrative condition. Some of this difference is likely

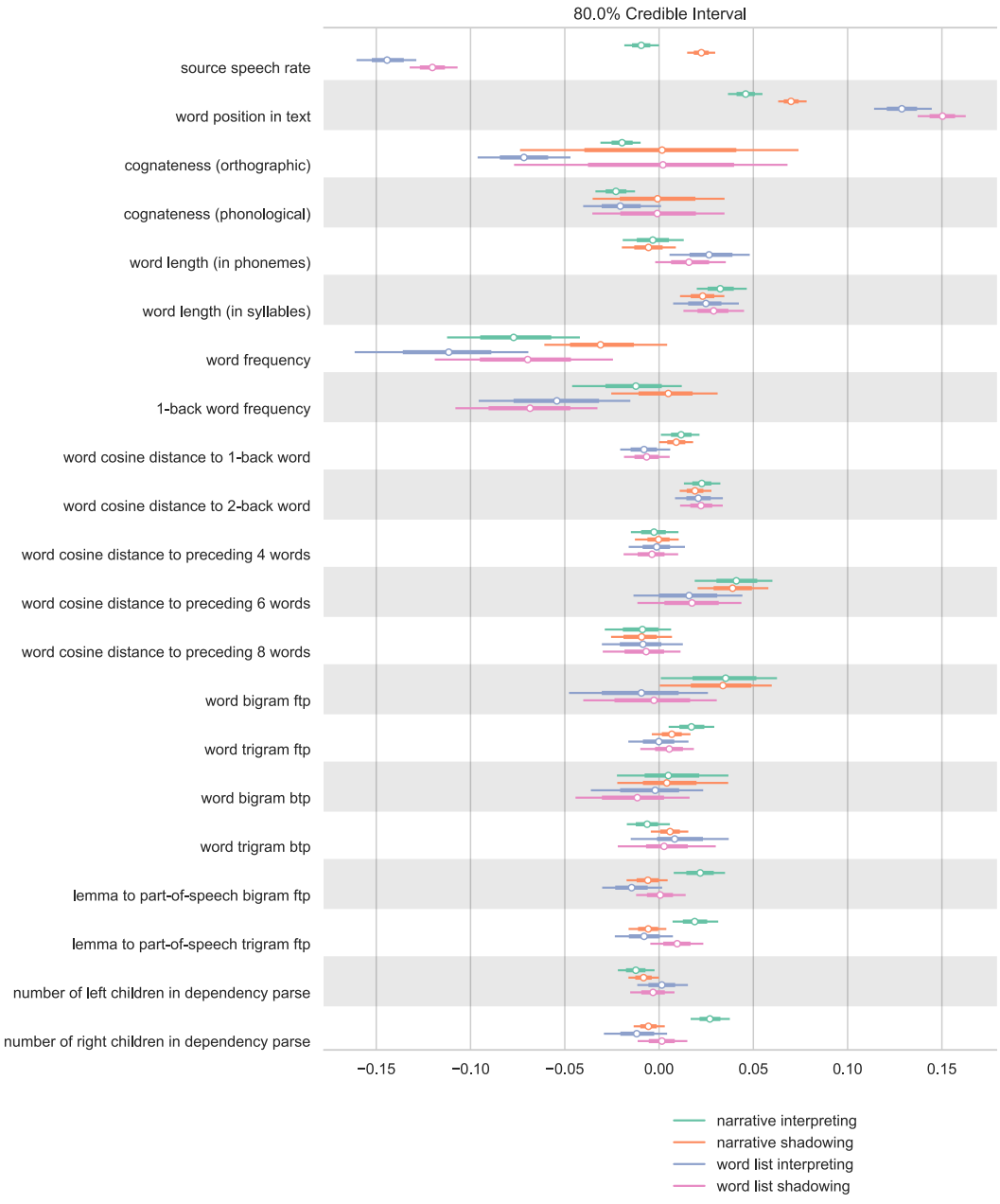


Figure 4.6: 80% Credible intervals of coefficient estimates from sparse Bayesian regression.

due to smaller number of observations in the latter condition, because the texts were shorter (due to the removal of function words) and participants made more errors, resulting in fewer usable speech latencies. The remainder of this difference was possibly caused by the smaller value ranges in some predictors, as a result of the lack of narrative structure (e.g., values in the BTP predictor were uniformly low in the randomized word list condition, because random sequences of content words have low transitional probability).

4.3.3 Speech rate and word position effects

The model contains two non-linguistic predictors: Speech rate and word position in text. We expected increased speech rate to result in faster production, not because of any facilitatory effect, but because the more rapid presentation of novel input necessitated faster reproduction. Curiously, while the expected effect occurred in the randomized word list condition, it was absent in the narrative interpreting condition, and even reversed in the narrative shadowing condition. Participants may have used latency strategically, keeping a number of words in working memory before reproducing them in order to allow for word order differences between languages and reformulation of sentences. However, if participants were strategically taking in a certain number of words, we would expect markedly lower latencies at higher speech rates, particularly for interpreting, which we do not observe. Instead, the latency seems to be task-dependent, with response latency on the randomized word lists seemingly reflecting a tradeoff between meeting a response deadline (the onset of the next word) that varies with speech rate, and accurate reproduction. In contrast, the (lack of a) relationship between response latencies and speech rates on the narratives likely reflects processing constraints of some kind; an optimal amount of input to keep in working memory or a perceptual buffer before shadowing or interpreting it.

Word position in text had a similarly differential effect between the randomized word list and narrative conditions, but all of the effects were in the expected (slowing) direction, possibly reflecting an accumulation of fatigue occurring over the course of each text. This effect was stronger in the randomized condition than in the narrative condition, which might seem unexpected given the lower real speech rate in the randomized word list condition, but it is consistent with the observation that error rates were higher for the randomized word lists than for narratives (Chapter 3).

4.3.4 Lexical effects

Uncertainty in the estimate effect size for cognateness is large in the shadowing conditions, because cognateness was invariant in the shadowing task (edit distances are all zero because the input and output phonological words are identical), but it has the expected facilitatory effect in the interpreting task, albeit split between the orthographic and phonological measures of cognateness in proportions that vary between narrative interpreting and randomized word list interpreting. These measures are only moderately correlated and they seem to capture different aspects of cognate status, however the differing effect sizes are possibly due to weak identifiability of the variance that they do share (i.e., there is variance in speech latencies that is explained by both cognateness predictors, but the model cannot decide which predictors to apportion this common explained variance to).

Word length in syllables has the expected slowing effect in all four conditions. Additionally, word length in phonemes has a slowing effect in the randomized word list interpreting condition, but appears to have a negligible effect on the other three conditions. Collectively there appears to be a general effect of word length, with number of syllables just being a more stable measure of length than number of phonemes.

Word frequency has a considerably larger effect on interpreting than in shadowing, likely because the facilitatory effect of word frequency compounds when translating. More frequent words are recognized and retrieved faster, both in the source and the target language. High frequency source words tend to be paired with high frequency target words, both because frequency is strongly intertwined with word meaning, and because English and Dutch are closely related languages. This effect is larger in the randomized word list conditions than in the narrative conditions, which we attribute to there being more non-zero contextual effects in the narrative conditions. Several of these contextual effects are moderately correlated with word frequency; it is plausible that shared explained variance is partially absorbed into these contextual effects, making the word frequency effect look smaller in the narrative conditions.

1-back word frequency appears to only affect word list speech latencies, possibly because the additive effect of the ease of reproducing the previous word is more easily discerned in discontinuous, word-for-word production than in production that is planned at the level of phrases or sentences (as in the narrative conditions).

4.3.5 Transitional probability effects

Forward transitional probability presents a complicated picture. Counterintuitively, the coefficient estimates for several of the FTP measures in the narrative interpreting condition are positive, indicating slower responses for high FTP words. When excluding all other lexical frequency and probability effects and regression only FTP on speech latencies, the FTP parameter flips and becomes negative as expected. The positive parameter in the full model can therefore be understood as a consequence of the mathematical equivalences described in Chapter 5. Similarly, backward transitional probability effects

likely shrank to zero because there is a highly correlated predictor that is more predictive of the observed data.

4.3.6 Semantic context effects

Semantic similarity to 1-back and 2-back words suggests an increasing facilitatory effect with time: The 1-back word has only a small facilitatory effect, and only in the narrative conditions. For the 2-back word though, there is a larger facilitatory effect, and it is visible in all conditions. That this effect increases with time suggests some sort of spreading activation mechanism that requires time before it is able to exert top-down facilitation (akin to SOA-dependent facilitation in a more controlled priming experiment).

Semantic similarity to preceding context had a facilitatory effect in both narrative interpreting and narrative shadowing for the larger 6-word sliding context window, while the 4-word and 8-word context windows effects are shrunk to zero. These predictors are strongly correlated, but the model selected the 6-word context window and it appears the residual variance does not support additional effects from the other two context windows. In the randomized word list condition, none of the context windows produced a semantic facilitation effect discernably different from zero, indicating that participants did not experience facilitation from semantic context in these word lists.

4.3.7 Syntactic context effects

In the narrative interpreting condition, the number of left syntactic children has a very small, but non-zero facilitatory effect, while the number of right syntactic children has a clear slowing effect. These effects are shrunk to zero in the shadowing condition. The parse trees created by the SpaCy dependency parser for the randomized word lists are non-interpretable, because there is no actual syntax in

the word lists. The effect of the number of left and right children is therefore zero in the randomized word list conditions, as expected.

4.3.8 Interim discussion

From a statistical perspective, it is worth considering that there are many ways to perform variable selection for regression models, but many suffer from the problem that due to collinearity and covariance, systems of predictors do not behave simply as the sum of those predictors. The benefit of using a sparsity-inducing method, rather than variable or model selection using p-values or likelihood ratios is that we can consider all predictors that could plausibly affect our observed data as a single system, and rely on the inference procedure to determine which predictors are actually irrelevant (Piironen & Vehtari, 2017a).

One drawback to this method is that it encourages a decision between collinear predictors that is potentially artificial: We know that word length and word frequency are highly correlated, but the sparsity priors will result in common explained variance in the observed data being apportioned to one of the two predictors, leaving the estimated coefficient deceptively small. In some cases this might be desirable, if for instance one predictor is causally dependent on another, we might want the latter to soak up most of the common explained variance (but depending on measurement noise, the opposite might happen). In other cases, two predictors might be related in a manner that is not easily identified or separated, for instance in the case of similarity to semantic context and forward transitional probability; these measures are weakly to moderately correlated, but it is difficult to decide which of these two measures should have "primacy" in language processing.

To compute a measure of "explanatory potential" of a predictor, we can perform a Principal Components Regression (PCR). We orthogonalize the complete matrix of predictors using Principal Components Analysis (PCA) and then per-

form the same sparse regression that we performed with the original predictor matrix. The resulting posterior estimates of the coefficients will not suffer from issues with apportioning variance because by orthogonalizing we have removed collinearities. We can then multiply the posterior samples with the PCA factor loadings to compute the potential coefficients for the original predictors, rather than the observed coefficients. Each potential coefficient estimate is a linear combination of the principal components coefficient estimates. An important caveat here is that due to the transformations used in the procedure, the potential coefficients are not proper regression coefficients in the sense that if the potential coefficients cannot be used in a regression model to make actual predictions. Rather, they reflect the potential that a predictor has to contribute to explaining observed variance.

4.3.9 Principal components regression results

Figure 4.7 shows the potential coefficient-estimates and their credible intervals from principal components regression. Most coefficient estimates are remarkably similar, which is perhaps unsurprising given that the sparsifying priors cause the original model to apportion explained variance to predictors that have a large unique contribution, similar to the principal components we compute in the first step of the principal components regression. The most obvious difference between the two models is the shifted intervals around some of the frequency and transitional probability effects. As in the original model, the lexical frequency effects are in the expected (facilitatory) direction, while the transitional probability effects are counter-intuitively in the opposite direction. The shifted uncertainty here reflects that there are large common variance components that are exchangeable, while many smaller variance components are shrunk to zero, with some degree of uncertainty around that value. In the original sparse regression, there is only one source of uncertainty around the zero, and a prior specifically

shrinking the estimates towards that value. In the PCR model, there are a number of small, sparsified components that contribute to a reconstructed potential coefficient estimate, which appears to shift the intervals somewhat, further underlining the weak identifiability of collinear frequency and transitional probability effects.

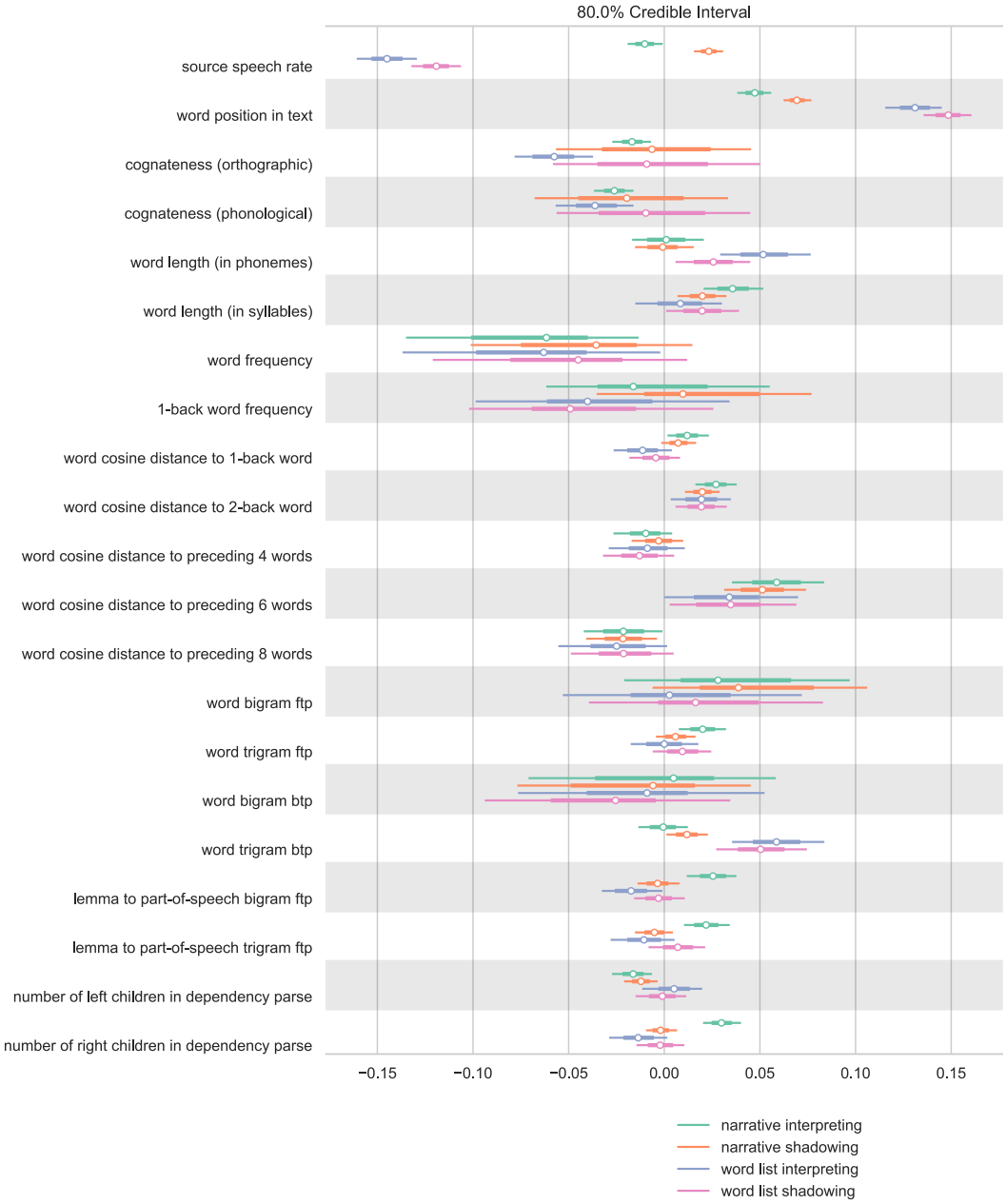


Figure 4.7: 80% Credible intervals of potential coefficient-estimates from principal components regression.

4.4 Discussion

Our lexical bottleneck model of speech processing in simultaneous interpreting and shadowing (Chapter 2) was found to not generalize well to speech without narrative structure, namely randomized word lists (Chapter 3). We hypothesized that the model fails to generalize because it does not include facilitation from narrative context. To assess this hypothesis, we investigated the nature and magnitude of contextual facilitation effects for narratives as opposed to randomized word lists in speech shadowing and simultaneous interpreting. We analyzed the speech latencies using Bayesian regression with sparsity-inducing priors to select relevant priors from a variety of lexical and contextual factors. The regression coefficients revealed differences between interpreting and shadowing in narratives and randomized word lists that indicate differences in planning scope and semantic processing. In the remainder of this section, we discuss our findings and their implications.

In the narrative shadowing condition, many of the lexical and contextual effects were weaker than in the narrative interpreting condition. This is not attributable to any underlying statistical phenomena; we therefore interpret this as reflecting real differences in linguistic processing between these two tasks.

Similarly, both shadowing and interpreting in the randomized word list condition exhibit lexical effects such as word length effects, but essentially no contextual effects. While the effect of context has both facilitatory and slowing aspects in our regression model, the higher error rates in the randomized word list condition suggest that removing narrative context overall has a negative effect (Chapter 3).

The absence of lexical frequency and word length effects in the narrative shadowing condition suggests that, when under considerable time pressure, participants do make use of a shallow, direct route from phonological input to phono-

logical output, bypassing the locus of these frequency effects. Similarly, the lack of syntactic effects in this condition suggests that syntactic processing does not occur in time to affect latency of production. That a facilitatory effect of semantic context still occurs (and for the same context window as in the interpreting condition) suggests that pre-activation at the semantic level does still occur. In contrast to the lexical effects, pre-activation from semantic context can build up over a longer time course and therefore facilitate even in the shadowing task. This is consistent with some participants self-reported conscious experience of performing the shadowing task: production follows perception, but precedes comprehension.

The length of the context window for which facilitation occurs suggests after 6 words pre-activation is too diffuse to be facilitatory, but that a shorter context window is not optimal either. That a shorter, more focal context window is suboptimal is somewhat surprising. It is unlikely that the semantic context effect reflects any kind of phrasal structure effect because the semantic vectors for the context window are simply averaged (cf. Pereira et al., 2018). It therefore does appear that the optimality of the 6-word context is a function of the nature of processing occurring at the semantic level during both interpreting and shadowing. The size of this context window is not, however, directly related to the latencies we observed, since the latencies in the narrative shadowing task ranged from 2 to 4 words (across the range of speech rates) but the model coefficients suggest an approximately 6-word context window for both shadowing and interpreting.

At the shorter 1-back and 2-word sliding context windows, there appears to be increasing facilitation in the narrative interpreting condition as time progresses past the presentation of a given word. The estimated effect sizes for these effects are very small, and we should therefore be careful to over-interpret the effects, but varying effects at different time scales have been reported in semantic prim-

ing studies using the continuous naming paradigm (see e.g., Rose & Rahman, 2017; Scaltritti et al., 2017, for discussion).

Because of the mathematical relationship between transitional probabilities and word frequencies, the counterintuitive slowing effect of FTP can be interpreted as a correction on facilitatory word frequency effects (cf. Chapter 5). However, it is unclear whether this also applies to the lemma to part-of-speech FTP, since the relationship between lemma to part-of-speech FTP is not fundamentally linked to word frequency the way regular FTP is. The exact nature of the slowing effect of lemma to part-of-speech FTP is an interesting avenue for further exploration.

The syntactic effects in narrative interpreting amount to facilitated production when a word closes syntactic dependencies that remained open, and a slowing when a word creates new syntactic dependencies. This is consistent with words being syntactically relatively more or less predictable from the preceding context. That neither of these effects occurs in narrative shadowing suggests that syntactic processing does not occur in time for it to affect the more direct processing route and relatively shorter latencies in the shadowing task. Alternately, it is possible that because the input-output latency is longer in interpreting, both in absolute (millisecond) terms and in terms of the number of words held in working memory, additional working memory load from tracking right dependencies builds up in interpreting, slowing down processing. This interpretation would appear to be at odds with the relatively long context being tracked at the semantic level, however it is possible that semantics and syntax are processed at different scopes (i.e., different context lengths).

The paradigms used in the present study were chosen because they allow us to elicit fairly consistent speech production across participants. Though in interpreting the specific sentence structure chosen by individual participants can vary, the high degree of similarity between English and Dutch and the task goal

of preserving the semantic content of the sentence results in participants producing similar sentences. This similarity in sentences produced by participants is crucial for making the statistical analyses reported in this study feasible, but unfortunately, it does come at the cost of making the linguistic content of the concurrent comprehension and production tasks highly similar (or even identical, in the case of shadowing) and therefore hard to disentangle. As an example, take similarity to semantic context: While we can compute this measure both for the input and output speech signals, those would be almost perfectly correlated, even in interpreting (due to the similarity between the two languages and the requirement to preserve the semantics of the input speech). The same applies to lexical factors such as word frequency or length, which are highly correlated between English and Dutch (and to each other, which further compounds the problem). In the present study we therefore do not attempt to model the effects of lexical and contextual factors on comprehension and production separately, instead opting to model them at the input speech level, knowing that the near-perfect correlation with the same factors for the output speech means we are in effect jointly modeling the effects of these factors on both comprehension and production.

In summary, these results show that while both interpreting and shadowing latencies are sensitive to lexical factors such as word length, contextual factors appear to affect shadowing and interpreting differentially. The absence of syntactic effects in shadowing reflects contextual processing in interpreting rather than in shadowing, and partly validates the modeling assumptions with regard to interpreting and shadowing routes made in Chapter 2. The low-level shadowing route hypothesized in that study allows words to be reproduced via connections at the phonological level, before they undergo full conceptual integration, meaning that comprehension in the full sense of the word can occur post-hoc (that is, after initiating articulatory planning or actual articulation) when shad-

owing. That semantic context nevertheless affects shadowing and interpreting equally indicates that semantic facilitation occurs during early processing, likely through top-down activation from semantic processing feeding into lower-level lexical or phonological processing. Most importantly, the differences between the narrative and randomized word list conditions make it clear that when modeling tasks such as shadowing or interpreting at the phrase or narrative level, it is crucial to account for continuous facilitation.

5 | A note on co-occurrence, transitional probability, and causal inference[1]

Abstract

Much has been written about the role of prediction in cognition in general, and language processing in particular, with some authors even positing that prediction is the central *goal* of cognition. Attributing a specific goal to cognition is speculative, but common theories of cognition posit that prediction plays a role in both perception and action. However, in studies on language processing, measures of predictability such as surprisal/forward transitional probability are no more, or even less effective in describing behavioral and neural phenomena than measures of post- or retrodictability such as backward transitional probability. We address this paradox by looking at the relationship between these different information theoretic measures and proposing a mechanistic account of how they are used in cognition. We posit that backward transitional probabilities support causal inferences about the occurrence of word sequences. Using Bayes' Theorem, we demonstrate that predictions (formalized as forward transitional probabilities) can be used in conjunction with the marginal probabilities of the current state/word and the upcoming state/word to compute these causal inferences. This conceptualization of causal inference in language processing both accounts for the role of prediction, and the surprising effectiveness of backwards transitional probability as a predictor of human behavior and its neural correlates.

[1] Adapted from Van Paridon, J. & Alday, P. M. (in prep.). A note on co-occurrence, transitional probability, and causal inference.

5.0.1 On *n*-gram frequency and conditional probability

For at least half a century, it has been recognized high frequency[2] words are easier to produce (Jescheniak & Levelt, 1994; Oldfield & Wingfield, 1965) and to recognize, both in speech (Broadbent, 1967; Cleland et al., 2006; Dahan et al., 2001) and in print (Cleland et al., 2006; Kuperman, 2013; Rayner, 1998). However, when modeling language processing (be it speech perception, reading, etc.), we are often interested in processing beyond the single word level. Processing at the level of multi-word phrases (word *n*-grams) is more complex to model than single-word processing. This is partly due to the (linear) increase in lexical factors when modeling multi-word phrases, but more importantly, our understanding of phrase processing is not as well developed as our understanding of single-word processing.

One easily accessible statistic relevant to phrase processing is word *n*-gram frequency, which has indeed been demonstrated to affect both language comprehension (Arnon & Snider, 2010) and language production (Janssen & Barber, 2012; Shao et al., 2019). These *n*-gram effects occur in addition to, and are distinct from, the effect of single-word frequency (Jacobs et al., 2016; Shao et al., 2019). It has been suggested that these *n*-grams are stored as single units (*lexical bundles*, see e.g., Jacobs et al., 2016; Tremblay et al., 2011). However, given the combinatorial explosion of word *n*-grams that occurs for any value of *n* greater than 1, it is clear that storing *n*-grams (in some sort of expanded mental lexicon) is infeasible for all but the highest frequency *n*-grams (Baayen et al., 2013), making whole *n*-gram storage inconsistent with the observation that *n*-gram frequency effects affect both high and low frequency *n*-grams (Arnon & Snider, 2010).

[2] We use the term frequency with regards to word occurrence in this article, which generally denotes a rate of occurrence (e.g., number of word occurrences per 1 million words, a scale from 0 to 1 million). Note however that for the purpose of comparing relative rates of occurrence, frequency is completely interchangeable with absolute counts (a scale from 0 to whatever the size of the corpus) and probabilities (a scale from 0 to 1).

We therefore reject the notion that apparent *n*-gram frequency effects are caused by the storage of whole *n*-grams and their frequencies, except for phrases with frequencies high enough to classify them as idioms (or compounds, cf. Jacobs & Dell, 2014), rather than phrases with purely compositional meaning. A more feasible mechanism than storing whole *n*-grams is to make use of conditional probabilities: The probability of a word occurring, given the occurrence of the preceding word. These conditional probabilities can be computed bidirectionally and are generally called transitional probabilities in the context of language (but note that these concepts are fundamentally equivalent). In studies of reading, the *forward transitional probability* is generally referred to as *predictability*, which has been found to have a robust effects on various reading-related measures (e.g., first fixation duration, Balota et al., 1985; and inspection probability, Kliegl et al., 2004; for an alternative implementation of predictability see McDonald and Shillcock, 2003). Transitional probabilities can also be reframed as *surprisal* ($-\log P_{\text{conditional}}$), an information theoretic measure that is often used in the field of Natural Language Processing. If we conceptualize the mental lexicon as a network of nodes and edges, transitional probabilities could feasibly be encoded in the edge weights, whereas storing whole *n*-grams requires an exponential increase in the number of nodes (cf. Baayen et al., 2013).

Hard to tell the difference: Equivalences

In the following sections, we transform all relevant quantities to a logarithmic scale (which is common practice) for reasons of computational convenience and cognitive plausibility[3].

Forward transitional probability (FTP) is a function of bigram and word$_1$ frequency:

$$\log P(w|w_{\text{prev}}) = \log \frac{P(w_{\text{prev}}, w)}{P(w_{\text{prev}})} = \log P(w_{\text{prev}}, w) - \log P(w_{\text{prev}}) \tag{5.1}$$

Backward transitional probability (BTP) is a function of bigram and word$_2$ frequency:

$$\log P(w_{\text{prev}}|w) = \log \frac{P(w_{\text{prev}}, w)}{P(w)} = \log P(w_{\text{prev}}, w) - \log P(w) \tag{5.2}$$

When we compile bigram occurrences from a large corpus of transcribed pseudoconversational speech (for corpus details see Chapter 6), we find a strong negative correlation between first- and second-word frequency (see Figure 5.1), as well as FTP and BTP (see Figure 5.2). Consequently, when first- and second-word frequency or FTP and BTP are both included in a linear model, the magnitude and direction of their effects will interact and therefore not be easily interpretable (if interpretable at all).

[3] Evidence suggests that word frequencies are experienced (both consciously and subconsciously) on a logarithmic scale. Contrast *angry* and *enraged*, for instance: *angry* is a fairly frequent word and *enraged* is fairly infrequent (in fact, in our dataset, *angry* is 56 times more frequent than *enraged*), however the effect of this difference in frequency on the difference in e.g., reading times or lexical decision times will not be proportional to the frequency, but to the logarithm of the difference in frequency. Similarly, when asked for explicit ratings in the difference in word frequency between different words, people are likely to give answers proportional to the logarithm of the frequency. This is in line with other power laws in cognition and perception and the reason why common measurement scales such as decibels for sound intensity are logarithmic in the physical unit, but linear in perception. Note also that the base of the logarithm is generally irrelevant, because every logarithm is a multiple of every other logarithm: When rescaling predictors to their standard deviation (common practice for linear regression in cognitive science), the rescaled predictor will be invariant with respect to the base of the logarithm because the standard deviation of a log-transformed predictor is proportional to the base of the logarithm.

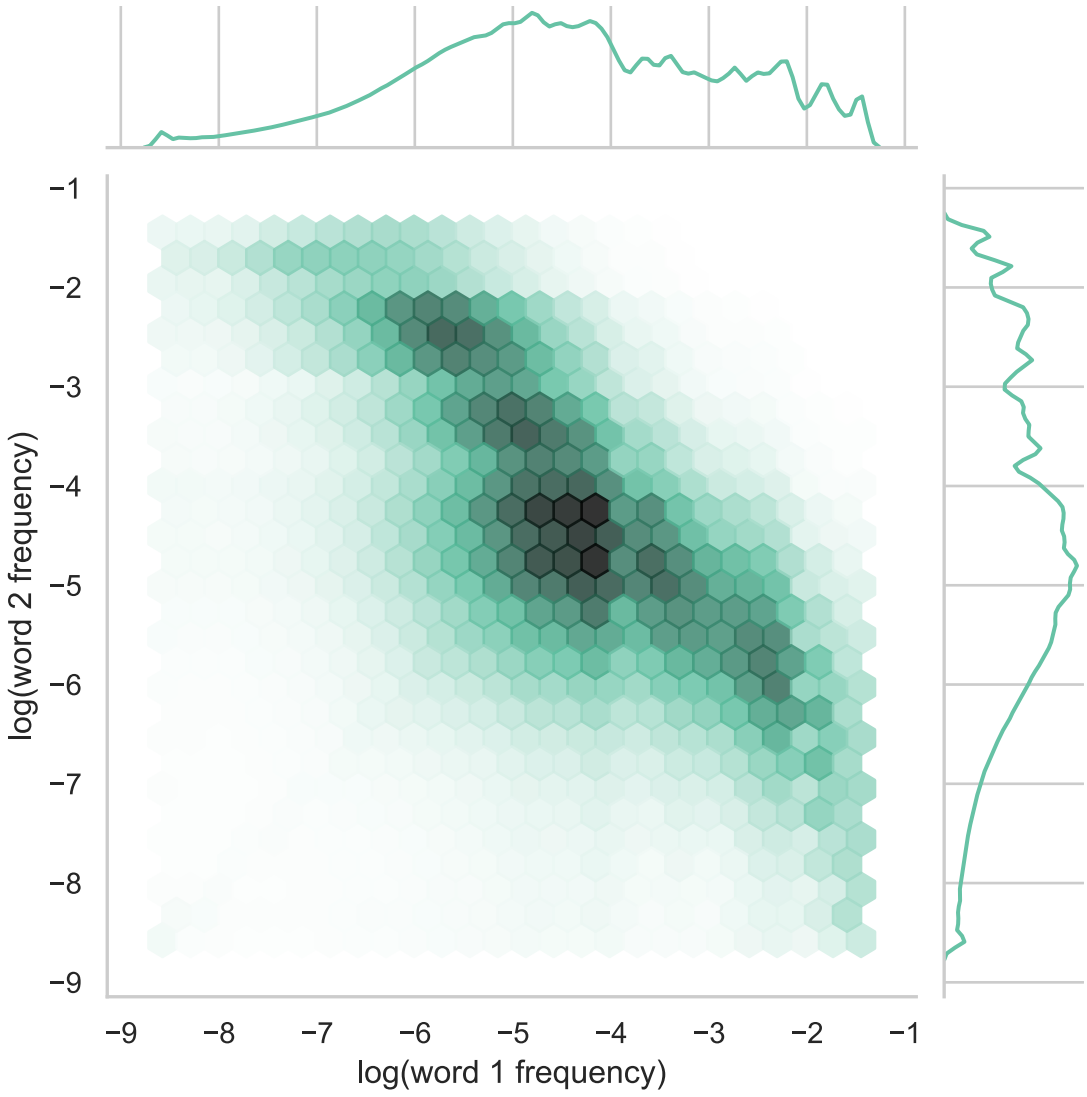


Figure 5.1: Joint distribution of first and last word frequencies for bigrams.

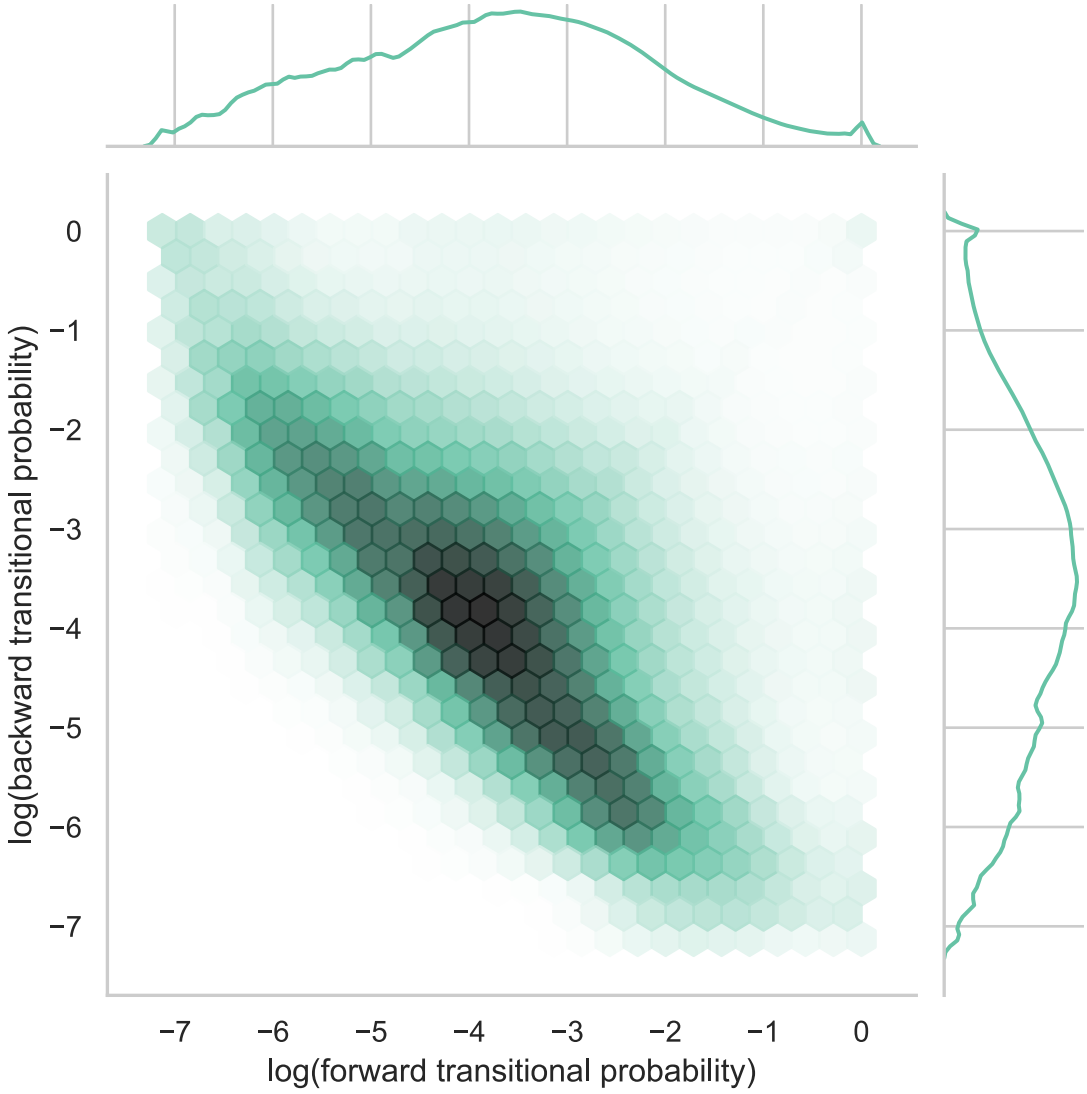


Figure 5.2: Joint distribution of forward and backward transitional probabilities for bigrams.

Unfortunately, things only get more confusable (and confusing) from here. Using Bayes' Theorem[4] we can compute FTP from BTP (and vice versa):

$$\log P(w|w_{\text{prev}}) = \log \frac{P(w_{\text{prev}}|w) \cdot P(w)}{P(w_{\text{prev}})} = \log P(w_{\text{prev}}|w) + \log P(w) - \log P(w_{\text{prev}}) \tag{5.3}$$

Ergo, information-theoretic surprisal, which is equivalent to negative log FTP can be computed from word frequencies and BTP[5]:

$$-\log P(w|w_{\text{prev}}) = \log P(w_{\text{prev}}) - \log P(w_{\text{prev}}|w) - \log P(w) \tag{5.6}$$

Similar results can be derived for other information theoretic measures.

The surprising effect of surprisal: Multicollinearity in linear models of behavior

The practical consequence of the equivalences outlined above is that when multiple measures of frequency and co-occurrence are used simultaneously as predictors in a linear model, this tends to result in collinearity between linear combinations of predictors. If this multicollinearity is perfect, it is impossible to perform the linear algebra necessary to fit the regression models. Most statistics packages will issue a warning regarding this multicollinearity and which predictors it con-

[4] Bayes' Theorem as used here is simply the law of conditional probabilities. Its use here is not specific to Bayesian statistics.

[5] Similarly, pointwise mutual information (PMI) can be computed from frequencies:

$$\log \frac{P(w_{\text{prev}}, w)}{P(w_{\text{prev}}) \cdot P(w)} = \log P(w_{\text{prev}}, w) - \log P(w) - \log P(w_{\text{prev}}) \tag{5.4}$$

And considering Equations 5.1 and 5.2, that means we can compute PMI from transitional probability (symmetrically):

$$\log \frac{P(w_{\text{prev}}, w)}{P(w_{\text{prev}}) \cdot P(w)} = \log P(w|w_{\text{prev}}) - \log P(w) = \log P(w_{\text{prev}}|w) - \log P(w_{\text{prev}}) \tag{5.5}$$

Many information-theoretical measures can be trivially computed from frequencies and transitional probabilities in this fashion.

cerns. However, even in cases where there is not perfect multicollinearity, the use of two or more co-occurrence measures can lead to unexpected consequences.

A hypothetical example: To predict reading times of a word of interest, w, we use w frequency and FTP from w_{prev} to w as predictors (the former as a measure for the ease of retrieving the current word, the latter as a measure for the predictability of the upcoming word). Counterintuitively, we find that low FTP is associated with *faster* reading. Does this mean that surprising words are somehow also more predictable? That seems contradictory. However, let's consider a simple linear model of reaction time with word frequency and FTP as predictors. For simplicity, we leave out the error term:

$$\log RT = \beta_0 + \beta_1 \cdot \log P(w) + \beta_2 \cdot \log P(w|w_{\text{prev}})$$

Now, using Equation 5.3, we note that:

$$\log P(w|w_{\text{prev}}) = \log P(w_{\text{prev}}|w) + \log P(w) - \log P(w_{\text{prev}})$$

And we substitute this back into the model:

$$\log RT = \beta_0 + \beta_1 \cdot \log P(w) + \beta_2 \cdot \big(\log P(w_{\text{prev}}|w) + \log P(w) - \log P(w_{\text{prev}})\big)$$

From this, it becomes clear that the apparent negative effect for FTP on actually reflects an effect of BTP minus w_{prev} frequency.

$$\log RT = \beta_0 + (\beta_1 + \beta_2) \cdot \log P(w) + \beta_2 \cdot \big(\log P(w_{\text{prev}}|w) - \log P(w_{\text{prev}})\big)$$

Because the effect of word frequencies is so large and stable, w_{prev} frequency will tend to dominate BTP and the coefficient β_2 will be negative. At the same

time, because β_2 is negative, β_1 will be inflated making the effect of w frequency seem larger than it is.

5.0.2 On making theoretically motivated choices

That certain combinations of frequency and transitional probability measures are mathematically exchangeable might seem convenient because it allows us to choose a set of predictors that is convenient for us to work with. This becomes problematic however if we try to claim that cognitive processes operate on the specific selection of quantities that we (arbitrarily) chose to model. Strong correlations and multicollinearity make it near impossible for naive statistical methods to distinguish between theoretical accounts that posit the importance of one probability measure (be it transitional probability, unigram or bigram frequency, etc.) over others (see e.g., Levy, 2008, for a footnote on how FTP/BTP correlation complicates predictor selection). Fundamentally, the psychological or neurobiological implementation of processes sensitive to frequency and transitional probability matters, but we are not able to draw conclusions from these measures alone. Instead, we need to look to fundamental structural reasons why one representation would be more compatible with other structures and mechanisms, such as temporal structure, causality, and basic insights regarding neural connectivity (much the same way that arguments about frequency versus time domain representation in M/EEG are resolved by proposing fundamental mechanisms and not by computing the Fourier transform).

Rethinking predictive coding: Retrodiction as inference

Rather than putting prediction central in cognition, we posit that cognition functions by making probabilistic causal inferences. If we start from the assumption that at its core, cognition subserves a perception-action loop, a particularly useful cognitive mechanism would be to compute inferences about the state of the

external world and the things that led to the current state (i.e., causality), as this is can guide both (imperfect) perception and subsequent action planning. Inferences about the current state of the world and the chain of states leading to the current state are encoded as backward transitional probabilities. The backward transitional probability directly answers the question "how probable is it, that the currently observed state was preceded by a given state?". This probabilistic notion of causality is the same type used in *Granger causality*: it does not imply causality in the philosophical or physical sense, but it does imply stochastic sequential dependence (Granger, 1969). This inference is computed via Bayes' Theorem, as above (Equation 5.3). In particular, we use information about marginal probabilities (of the current state (marginal likelihood) and the next state (prior)) combined with the conditional probability of the next state based on the current state (likelihood, here FTP) to compute probabilistic causality. Note that prediction occurs here as an intermediate step in determining causality: the likelihood, i.e., FTP, is a critical piece in computing causality. Note that this account also explains the relative success of measures such as cloze probability. In this framework, cloze probability corresponds to the maximum likelihood. In a typical experiment, where word frequencies have been carefully controlled, we thus have a manipulation of the likelihood under nearly constant priors. Because the maximum likelihood under constant priors is proportional to the maximum a posteriori value (MAP), i.e., the peak of the posterior, the standard cloze manipulation corresponds to a manipulation of the BTP. At the same time, we do not have perfect control over the prior (word frequency) in experiments, and so the maximum likelihood does not directly correspond to the MAP.

Bayesian brains with *some* probabilities?

Although we have presented our account as a direct computation of probabilities, this is not a necessity. Indeed, our account is compatible with sampling

perspectives with or without direct knowledge of probabilities (cf. Sanborn & Chater, 2016) and with variational accounts (Friston, 2005; Friston et al., 2012). It is consistent with the "reversal" of the flow of prediction and error in prominent accounts such as Friston's (2005) theory of cortical responses. In this theory, prediction flows upward through the cortical hierarchy, while error propagates downward. In our account, prediction is used to compute the probability of a given cause, which corresponds to the goodness of fit, or equivalently error, associated with that cause.

5.0.3 Conclusion

The rise of information theory in the brain and behavioral sciences has presented researchers with a plethora of potential quantitative measures. We posit that the (arbitrary) choice of measure cannot be driven by purely statistical concerns, because commonly used measures are linear combinations of each other and thus statistically indistinguishable. This complex interrelation gives rise to apparent paradoxes, such as an illusory facilitation in processing surprising words when controlling for absolute frequency. However, these paradoxes should not be overinterpreted, as they are spurious, introduced by a particular decomposition. Instead, we should focus on computational accounts, such as the one proposed here. By assuming that inferences about causality are instrumental in both perception and action, two of the core operations of cognition, we arrive at an account of prediction as a side effect, rather than a "goal" of cognition. This account allows us to make theoretically motivated choices between information theoretic measures as predictors for language processing and human behavior more generally.

6 | subs2vec: Word embeddings from subtitles in 55 languages[1]

Abstract

This paper introduces a novel collection of word embeddings, numerical representations of lexical semantics, in 55 languages, trained on a large corpus of pseudo-conversational speech transcriptions from television shows and movies. The embeddings were trained on the OpenSubtitles corpus using the fastText implementation of the skipgram algorithm. Performance comparable with (and in some cases exceeding) embeddings trained on non-conversational (Wikipedia) text is reported on standard benchmark evaluation datasets. A novel evaluation method of particular relevance to psycholinguists is also introduced: prediction of experimental lexical norms in multiple languages. The models, as well as code for reproducing the models and all analyses reported in this paper (implemented as a user-friendly Python package), are freely available at: https://github.com/jvparidon/subs2vec.

[1] Adapted from Van Paridon, J. & Thompson, B. (2020). subs2vec: Word embeddings from subtitles in 55 languages. *Behavior Research Methods*. doi:10.3758/s13428-020-01406-3.

6.1 Introduction

Recent progress in applied machine learning has resulted in new methods for efficient induction of high-quality numerical representations of lexical semantics – *word vectors* – directly from text. These models implicitly learn a vector space representation of lexical relationships from co-occurrence statistics embodied in large volumes of naturally occurring text. Vector representations of semantics are of value to the language sciences in numerous ways: as hypotheses about the structure of human semantic representations (e.g., Chen et al., 2017); as tools to help researchers interpret behavioral (e.g., Pereira et al., 2016) and neurophysiological data (e.g., Pereira et al., 2018), and to predict human lexical judgements of e.g., word similarity, analogy, and concreteness (see Methods for more detail); and as models that help researchers gain quantitative traction on large-scale linguistic phenomena, such as semantic typology (e.g., Thompson et al., 2018), semantic change (e.g., Hamilton et al., 2016), or linguistic representations of social biases (e.g., Garg et al., 2018), to give just a few examples.

Progress in these areas is rapid, but nonetheless constrained by the availability of high quality training corpora and evaluation metrics in multiple languages. To meet this need for large, multilingual training corpora, word embeddings are often trained on Wikipedia, sometimes supplemented with other text scraped from web pages. This has produced steady improvements in embedding quality across the many languages in which Wikipedia is available (see e.g., Al-Rfou et al., 2013; Bojanowski et al., 2017; Grave et al., 2018)[2]; large written corpora meant as repositories of knowledge. This has the benefit that even obscure words and semantic relationships are often relatively well-attested.

However, from a psychological perspective, these corpora may not represent the kind of linguistic experience from which people learn a language, raising con-

[2]More examples can be found in this Python package that collects recent word embeddings: https://github.com/plasticityai/magnitude

cerns about psychological validity. The linguistic experience over the lifetime of the average person typically does not include extensive reading of encyclopedias. While word embedding algorithms do not necessarily reflect human learning of lexical semantics in a mechanistic sense, the semantic representations induced by any effective (human or machine) learning process should ultimately reflect the latent semantic structure of the corpus it was learned from.

In many research contexts, a more appropriate training corpus would be one based on conversational data of the sort that represents the majority of daily linguistic experience. However, since transcribing conversational speech is labor-intensive, corpora of real conversation transcripts are generally too small to yield high quality word embeddings. Therefore, instead of actual conversation transcripts, we used television and film subtitles since these are available in large quantities.

That subtitles are a more valid representation of linguistic experience, and thus a better source of distributional statistics, was first suggested by New et al. (2007) who used a subtitle corpus to estimate word frequencies. Such subtitle-derived word frequencies have since been demonstrated to have better predictive validity for human behavior (e.g., lexical decision times) than word frequencies derived from various other sources (e.g., the Google Books corpus and others; Brysbaert et al., 2011; Brysbaert & New, 2009; Keuleers et al., 2010). The SUBTLEX word frequencies use the same OpenSubtitles corpus used in the present study. Mandera et al. (2017) have previously used this subtitle corpus to train word embeddings in English and Dutch, arguing that the reasons for using subtitle corpora also apply to distributional semantics.

While film and television speech could be considered only pseudo-conversational in that it is often scripted and does not contain many disfluencies and other markers of natural speech, the semantic content of TV and movie subtitles better reflects the semantic content of natural speech than the commonly used corpora

of Wikipedia articles or newspaper articles. Additionally, the current volume of television viewing makes it likely that for many people, television viewing represents a plurality or even the majority of their daily linguistic experience. For example, one study of 107 preschoolers found they watched an average of almost three hours of television per day, and were exposed to an additional four hours of background television per day (Nathanson et al., 2014).

Ultimately, regardless of whether subtitle-based embeddings outperform embeddings from other corpora on the standard evaluation benchmarks, there is a deeply principled reason to to pursue conversational embeddings: The semantic representations learnable from *spoken* language are of independent interest to researchers studying the relationship between language and semantic knowledge (see e.g., Lewis et al., 2019; Ostarek et al., 2019).

In this paper we present new, freely available, subtitle-based pretrained word embeddings in 55 languages. These embeddings were trained using the fastText implementation of the skipgram algorithm on language-specific subsets of the OpenSubtitles corpus. We trained these embeddings with two objectives in mind: to make available a set of embeddings trained on transcribed pseudo-conversational language, rather than written language; and to do so in as many languages as possible to facilitate research in less-studied languages. In addition to previously published evaluation datasets, we created and compiled additional resources in an attempt to improve our ability to evaluate embeddings in languages beyond English.

6.2 Method

6.2.1 Training corpus

To train the word vectors, we used a corpus based on the complete subtitle archive of OpenSubtitles.org, a website that provides free access to subtitles

contributed by its users. The OpenSubtitles corpus has been used in prior work to derive word vectors for a more limited set of languages (only English and Dutch; Mandera et al., 2017). Mandera and colleagues compared skipgram and CBOW algorithms as implemented in word2vec (Mikolov, Chen, et al., 2013) and concluded that when parameterized correctly, these methods outperform older, count-based distributional models. In addition to the methodological findings, Mandera and colleagues also demonstrated the general validity of using the OpenSubtitles corpus to train word embeddings that are predictive of behavioral measures. This is consistent with the finding that the word frequencies (another distributional measure) in the OpenSubtitles corpus correlate better with human behavioral measures than frequencies from other corpora (Brysbaert et al., 2011; Brysbaert & New, 2009; Keuleers et al., 2010).

The OpenSubtitles archive contains subtitles in many languages, but not all languages have equal numbers of subtitles available. This is partly due to differences in size between communities in which a language is used and partly due to differences in the prevalence of subtitled media in a community (e.g., English language shows broadcast on Dutch television would often be subtitled, whereas the same shows may often be dubbed in French for French television). While training word vectors on a very small corpus will likely result in impoverished (inaccurate) word representations, it is difficult to quantify the quality of these vectors, because standardized metrics of word vector quality exist for only a few (mostly Western European) languages. We are publishing word vectors for every language we have a training corpus for, regardless of corpus size, alongside explicit mention of corpus size. These corpus sizes should not be taken as a direct measure of quality, but word vectors trained on a small corpus should be treated with caution.

6.2.2 Preprocessing

We stripped the subtitle and Wikipedia corpora of non-linguistic content such as time-stamps and XML tags. Paragraphs of text were broken into separate lines for each sentence and all punctuation was removed. All languages included in this study are space-delimited, therefore further parsing or tokenization was not performed. The complete training and analysis pipeline is unicode-based, hence non-ASCII characters and diacritical marks were preserved.

After preprocessing, we deduplicated the corpora in order to systematically remove over-represented, duplicate material from the corpus. While Mandera et al. (2017) deduplicated by algorithmically identifying and removing duplicate and near-duplicate subtitle documents, we performed deduplication by identifying and removing duplicate lines across the whole corpus for each language as advocated by Mikolov et al. (2017). This method was used for both the subtitle and Wikipedia corpora. Line-wise deduplication preserves different translations of the same sentence across different versions of subtitles for the same movie, thus preserving informative variation in the training corpus while still removing uninformative duplicates of highly frequent lines such as “Thank you!”.

Finally, bigrams with a high mutual information criterion were transformed into single tokens with an underscore (e.g., “New York” becomes “New_York”) in five iterations using the Word2Phrase tool with a decreasing mutual information threshold and a probability of 50% per token on each iteration (Mikolov, Sutskever, et al., 2013).

6.2.3 fastText skipgram

The word embeddings were trained using fastText, a collection of algorithms for training word embeddings via context prediction. FastText comes with two algorithms, CBOW and skipgram (see Bojanowski et al., 2017, for review).

Table 6.1: fastText skipgram parameter settings used in the present study.

Parameter	Value	Description
minCount	5	Min. number of word occurrences
minn	3	Min. length of subword ngram
maxn	6	Max. length of subword ngram
t	.0001	Sampling threshold
lr	.05	Learning rate
lrUpdateRate	100	Rate of updating the learning rate
dim	300	Dimensions
ws	5	Size of the context window
epoch	10	Number of epochs
neg	10	Number of negatives sampled in the loss function

A recent advancement in the CBOW algorithm, using position-dependent weight vectors, appears to yield better embeddings than currently possible with skipgram (Mikolov et al., 2017). No working implementation of CBOW with position-dependent context weight vectors has yet been published. Therefore, our models were trained using the current publicly available state of the art by applying the improvements in fastText parametrization described in Grave et al. (2018) to the default parametrization of fastText skipgram described in Bojanowski et al. (2017); the resulting parameter settings are reported in Table 6.1.

6.2.4 Evaluation of embeddings

A consensus has emerged around evaluating word vectors on two tasks: predicting human semantic similarity ratings and solving word analogies. In the analogies domain the set of analogies published by Mikolov, Sutskever, et al. (2013) has emerged as a standard and has been translated into French, Polish, and Hindi by Grave et al. (2018) and additionally into German, Italian, and Portuguese (Berardi et al., 2015; Köper et al., 2015; Querido et al., 2017). Semantic similarity ratings are available for many languages and domains (nouns, verbs, common words, rare words) but the most useful for evaluating relative success of word vectors

in different languages are similarity sets that have been translated into multiple languages: RG65 in English (Rubenstein & Goodenough, 1965), Dutch (Postma & Vossen, 2014), German (Gurevych, 2005) and French (Joubarne & Inkpen, 2011), MC30 (a subset of RG65) in English (Miller & Charles, 1991), Dutch (Postma & Vossen, 2014), and Arabic, Romanian, and Spanish (Hassan & Mihalcea, 2009), YP130 in English (Yang & Powers, 2006) and German (Meyer & Gurevych, 2012), SimLex999 in English (Hill et al., 2014) and Portuguese (Querido et al., 2017), Stanford Rare Words in English (Luong et al., 2013) and Portuguese (Querido et al., 2017), and WordSim353 in English (Finkelstein et al., 2001), Portuguese (Querido et al., 2017), and Arabic, Romanian, and Spanish (Hassan & Mihalcea, 2009).

Additional similarity datasets we could only obtain in just a single language are MEN3000 (Bruni et al., 2012), MTurk287 (Radinsky et al., 2011), MTurk771 (Halawi et al., 2012), REL122 (Szumlanski et al., 2013), SimVerb3500 (Gerz et al., 2016) and Verb143 (Baker et al., 2014) in English, Schm280 (a subset of WS353; Schmidt et al., 2011) and ZG222 in German (Zesch & Gurevych, 2006), FinnSim300 in Finnish (Venekoski & Vankka, 2017), and HJ398 in Russian (Panchenko et al., 2016).

Solving analogies

To add to the publicly available translations of the so-called Google analogies introduced by Mikolov, Chen, et al. (2013), we translated these analogies from English into Dutch, Greek, and Hebrew. Each translation was performed by a native speaker of the target language with native-level English proficiency. Certain categories of syntactic analogies are trivial when translated (e.g., adjective and adverb are identical wordforms in Dutch). These categories were omitted. In the semantic analogies, we omitted analogies related to geographic knowledge (e.g., country and currency, city and state) because many of the words in these analo-

gies are not attested in the OpenSubtitles corpus. Solving of the analogies was performed using the cosine multiplicative method for word vector arithmetic described by Levy and Goldberg (2014) (see Eq. 6.1).

$$\arg\max_{b^* \in V} = \frac{\cos(b^*, b)\cos(b^*, a^*)}{\cos(b^*, a) + \varepsilon} \tag{6.1}$$

For analogies of the form a is to a^* as b is to b^*. With small but non-zero ε to prevent division by zero. Equation reproduced here from Levy and Goldberg (2014).

Predicting lexical norms

To support experimental work, psycholinguists have collected large sets of *lexical norms*. Brysbaert, Warriner, et al. (2014), for instance, collected lexical norms of *concreteness* for 40,000 English words, positioning each on a 5-point scale from highly abstract to highly concrete. Lexical norms have been collected for English words in a range of semantic dimensions. Significant attention has been paid to *valence, arousal, dominance* (13K words, Warriner et al., 2013), and *age of acquisition* (30K words, Kuperman et al., 2012). Other norm sets characterize highly salient dimensions such as *tabooness* (Janschewitz, 2008). In a similar, but more structured study, Binder et al. (2016) collected ratings for 62 basic conceptual dimensions (e.g., *time, harm, surprise, loud, head, smell*), effectively constructing 62-dimensional psychological word embeddings that have been shown to correlate well with brain activity.

Norms have been collected in other languages too. Although our survey is undoubtedly incomplete, we collated published norm sets for various other, less studied languages (see Tables 6.2 and 6.3 for an overview). These data can be used to evaluate the validity of computationally induced word embeddings in multiple languages. Prior work has demonstrated that well-attested lexical norms (i.e., Valence, Arousal, Dominance, and Concreteness in English) can be predicted with reasonable accuracy using a simple linear transformation of word

embeddings (Hollis & Westbury, 2016). Using this approach, the lexical norms can be understood as gold-standard unidimensional embeddings with respect to human-interpretable semantic dimensions. In general this relationship has been exploited to use word embeddings to predict lexical norms for words that no norms are available for (e.g., Bestgen, 2008; Bestgen & Vincze, 2012; Dos Santos et al., 2017; Feng et al., 2011; Hollis et al., 2017; Recchia & Louwerse, 2015a, 2015b; Turney & Littman, 2002, 2003; Vankrunkelsven et al., 2015; Westbury et al., 2013), although this procedure should be used with caution, as it can introduce artefacts in a predicted lexical norm, especially for norms that are only weakly predictable from word embeddings (see Mandera et al., 2015, for an extensive discussion of this issue).

Conversely, the same relationship can be used as an evaluation metric for word embeddings by seeing how well new vectors predict lexical norms. Patterns of variation in prediction can also be illuminating: are there semantic norms that are predicted well by vectors trained on one corpus but not another, for example? We examined this question by using L2-penalized regression to predict lexical norms from raw word vectors. Using regularized regression reduces the risk of overfitting for models like the ones used to predict lexical norms here, with a large number of predictors (the 300 dimensions of the word vectors) and relatively few observations. Ideally, the regularization parameter is tuned to the amount of observations for each lexical norm, with stronger regularization for smaller datasets. However, in the interest of comparability and reproducibility, we kept the regularization strength constant. We fit independent regressions to each lexical norm, using five-fold cross validation repeated ten times (with random splits each time). We report the mean correlation between the observed norms and the predictions generated by the regression model, adjusted (penalized) for any words missing from our embeddings. Because of the utility of lexical norm prediction and extension (predicting lexical norms for unattested words),

Table 6.2: Lexical norms datasets. 1/3

Language	Article	Lexical norms	Number of words	Number of raters
Dutch	Brysbaert, Stevens, et al. (2014)	Age of acquisition, concreteness	25888	15 per item
Dutch	Keuleers et al. (2015)	Prevalence	52847	300 per item
Dutch	Roest et al. (2018)	Arousal, insulting, taboo (general), taboo (personal), valence	672	87 per item
Dutch	Speed and Majid (2017)	Arousal, auditory, dominance, gustatory, modality exclusivity, olfactory, tactile, valence, visual	485	15 per item
Dutch	Verheyen et al. (2019)	Age of acquisition, arousal, concreteness, dominance, familiarity, imageability, valence	1000	20 per item
English	Brysbaert, Warriner, et al. (2014)	Concreteness	37058	25 per item
English	Brysbaert et al. (2019)	Prevalence	61855	388 per item
English	Engelthaler and Hills (2018)	Humorousness	4997	35 per item
English	Janschewitz (2008)	Familiarity, offensiveness, tabooness, personal use	460	78 per item
English	Keuleers et al. (2012)	Lexical decision time	28515	39 per item
English	Kuperman et al. (2012)	Age of acquisition	30121	20 per item
English	Lynott et al. (2019)	Lancaster sensorimotor norms	39707	25 per item
English	Pexman et al. (2019)	Body-object interaction	9349	26 per item
English	Scott et al. (2019)	Age of acquisition, arousal, concreteness, dominance, familiarity, gender association, imageability, semantic size, valence	5553	20 per item
English	Warriner et al. (2013)	Arousal, dominance, valence	13915	20 per item

Table 6.3: Lexical norms datasets. 2/3

Language	Article	Lexical norms	Number of words	Number of raters
Farsi	Bakhtiar and Weekes (2015)	Age of acquisition, familiarity, imageability	871	40 per item
Finnish	Eilola and Havelka (2010)	Concreteness, emotional charge, familiarity, offensiveness, valence	210	150 per item
Finnish	Söderholm et al. (2013)	Arousal, valence	420	250 per item
French	Bonin et al. (2018)	Arousal, concreteness, context availability, valence	1659	30 per item
French	Chedid, Wilson, et al. (2019)	Familiarity	3596	20 per item
French	Chedid, Brambati, et al. (2019)	Auditory perceptual strength, visual perceptual strength	3596	25 per item
French	Desrochers and Thompson (2009)	Imageability	3600	72 per item
French	Ferrand et al. (2010)	Lexical decision time	38840	25 per item
French	Monnier and Syssau (2014)	Arousal, valence	1031	37 per item
German	Grandy et al. (2020)	Imageability, emotionality (in two age groups)	2592	20 per item
German	Kanske and Kotz (2010)	Arousal, concreteness, valence	1000	64 per item
German	Schauenburg et al. (2015)	Arousal, authority, community, potency, valence	858	35 per item
Indonesian	Sianipar et al. (2016)	Arousal, concreteness, dominance, predictability, valence	1490	70 per item

Table 6.4: Lexical norms datasets. 3/3

Language	Article	Lexical norms	Number of words	Number of raters
Italian	Vergallito et al. (2020)	Auditory, gustatory, haptic, lexical decision time, modality exclusivity, naming time, olfactory, visual	1121	57 per item
Malay	Yap et al. (2010)	Lexical decision time	1510	44 per item
Polish	Imbir (2015)	Arousal, concreteness, dominance, imageability valence	4905	50 per item
Portuguese	Cameirão and Vicente (2010)	Age of acquisition	1749	48 per item
Portuguese	Soares et al. (2012)	Arousal, dominance, valence	1034	50 per item
Spanish	Abella and González-Nosti (2019)	Age of acquisition, motor content	4565	25 per item
Spanish	Díez-Álamo et al. (2018)	Color vividness, graspability, pleasant taste, risk of pain, smell intensity, sound intensity, visual motion	750	26 per item
Spanish	Díez-Álamo et al. (2019)	Sensory experience	5500	35 per item
Spanish	Guasch et al. (2016)	Arousal, concreteness, context availability, familiarity, imageability, valence	1400	20 per item
Spanish	Stadthagen-Gonzalez et al. (2017)	Arousal, valence	14031	20 per item
Spanish	Stadthagen-González et al. (2018)	Anger, arousal, disgust, fear, happiness, sadness, valence	10491	20 per item
Turkish	Göz et al. (2017)	Age of acquisition, imagery, concreteness	600	457 per item

we have included a lexical norm prediction/extension module and usage instructions in the *subs2vec* Python package.

6.3 Results

Results presented in this section juxtapose three models generated by the authors using the same parametrization of the fastText skipgram algorithm: A *wiki* model trained on a corpus of Wikipedia articles, a *subs* model trained on the OpenSubtitles corpus, and a *wiki+subs* model trained on a combination of both corpora. A priori, we expected the models trained on the largest corpus in each language (wiki+subs) to exhibit the best performance. Performance measures are penalized for missing word vectors. For example: If for only 80% of the problems in an evaluation task word vectors were actually available in the subs vectors, but those problems were solved with 100% accuracy, the reported score would be only 80%, rather than 100%. If the wiki vectors on that same task included 100% of the word vectors, but only 90% accuracy was attained, the adjusted scores (80% vs 90%) would reflect that the Wikipedia vectors performed better. (Unpenalized scores are included in Appendix 6.C, for comparison.)

6.3.1 Semantic dissimilarities

Spearman's rank correlation between predicted similarity (cosine distance between word vectors) and human-rated similarity is presented in Figure 6.1. Performance is largely similar, even for datasets like the Stanford Rare Words dataset where the Wikipedia corpus, by virtue of being an encyclopedia, tends to have more and better training samples for these rare words.

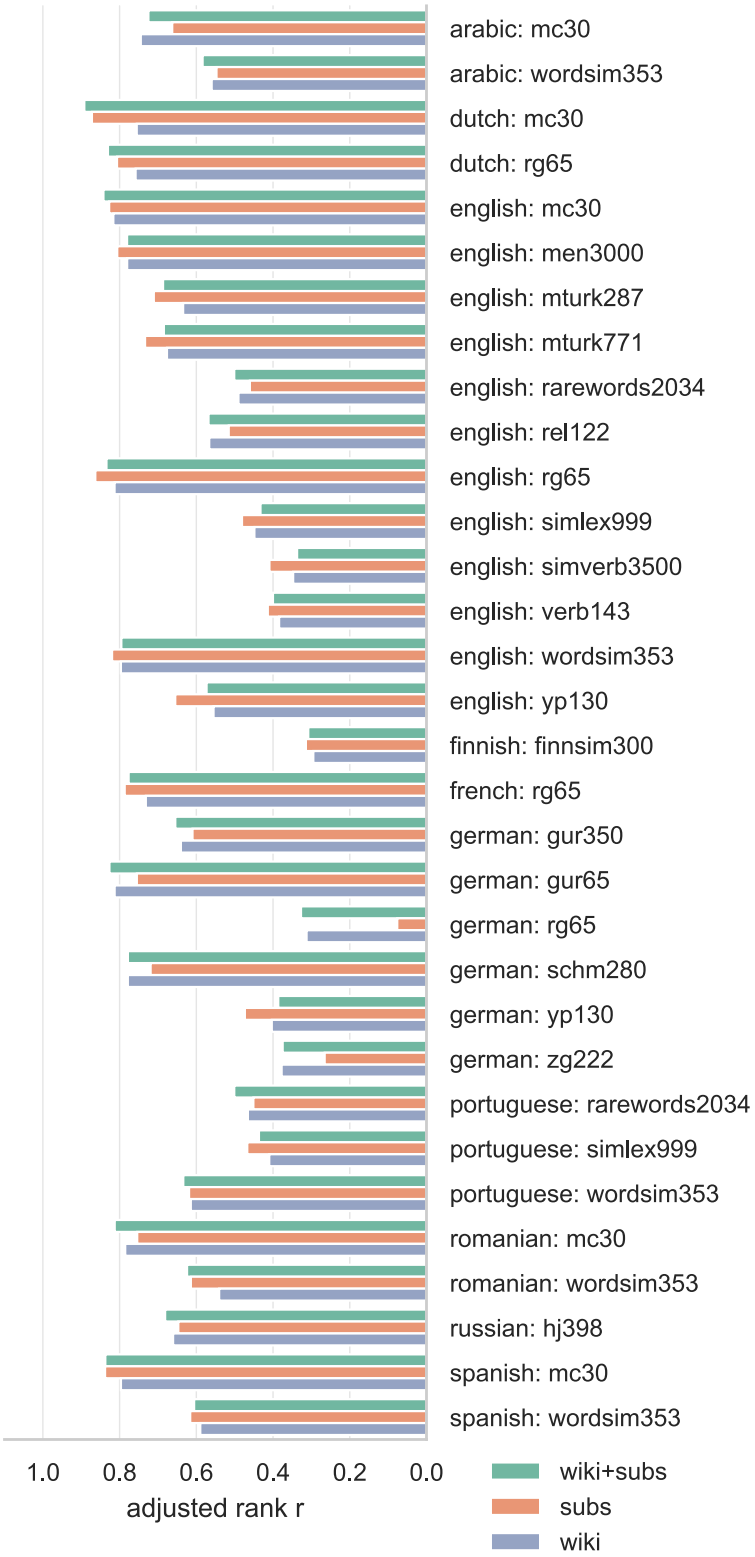


Figure 6.1: Rank correlations between human ratings of semantic similarity and word vector cosine similarity. Correlations are adjusted by penalizing for missing word vectors.

6.3.2 Semantic and syntactic analogies

Adjusted proportion of correctly solved analogies is presented in Figure 6.2. Note that while word vectors trained on a Wikipedia corpus strongly outperform the subtitle vectors on the semantic analogies sets, this is mostly due to a quirk of the composition of the semantic analogies: Geographic relationships of the type country-capital, city-state, or country-currency make up 93% of the commonly used semantic analogies. This focus on geographic information suits the Wikipedia-trained vectors, because being an encyclopedia, capturing this type of information is the explicit goal of Wikipedia. However, some of the more obscure analogies in this set (e.g., "Macedonia" is to "denar" as "Armenia" is to "dram") seem unlikely to be solvable for the average person (i.e., they do not appear to reflect common world knowledge). In this sense the lower scores obtained with the embeddings trained on the subtitle corpus are perhaps a better reflection of the linguistic experience accumulated by the average person. To better reflect general semantic knowledge, rather than highly specific geographic knowledge, we have removed the geographic analogies in the sets of analogies that were translated into new languages for the present study.

6.3.3 Lexical norms

Figures 6.3, 6.4, 6.5, and 6.6 show the adjusted correlation between observed lexical norms and the norms predicted by the word embedding models. Predictive accuracy for models trained on Wikipedia and OpenSubtitles is largely similar, with a notable exception for tabooness and offensiveness, where the models trained on subtitle data perform markedly better. Offensive and taboo words are likely not represented in their usual context on Wikipedia, resulting in word vectors that do not represent the way these words are generally experienced. The subtitle vectors, while not trained on actual conversational data, capture the con-

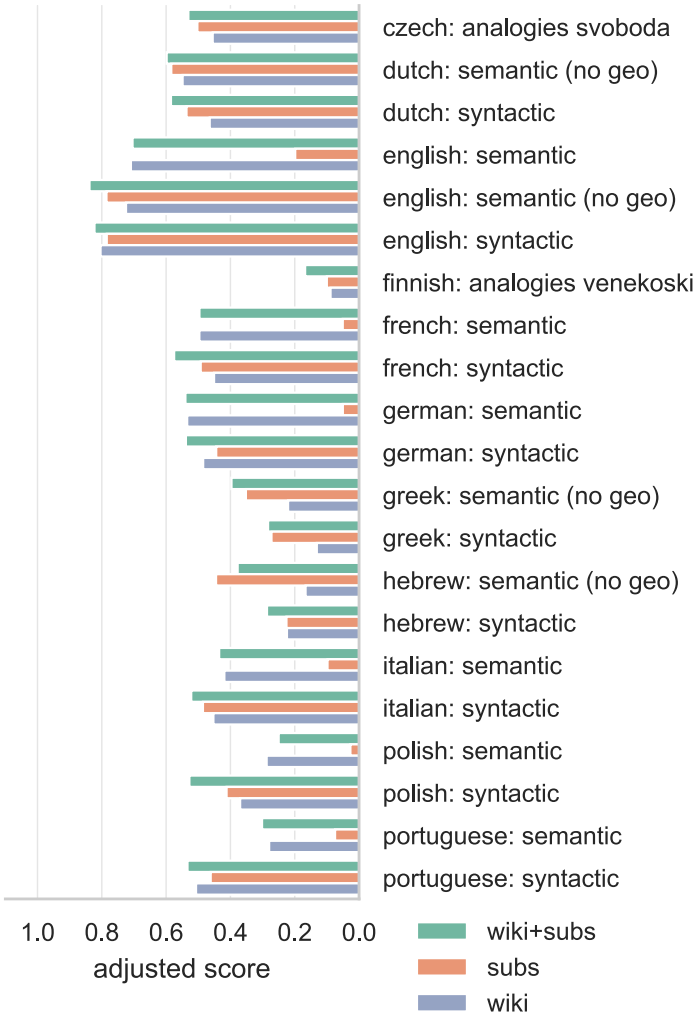


Figure 6.2: Proportion of correctly solved analogies in the semantic and syntactic domain using word vectors. Semantic datasets contained 93% geographic analogies, *no geo* datasets are those same datasets, excluding the geographic analogies. Scores are adjusted by penalizing for missing word vectors.

text in which taboo and offensive words are used much better. Models trained on a combined Wikipedia and OpenSubtitles corpus generally perform marginally better than either corpus taken separately, as predicted.

Figures 6.7 and 6.8 show the adjusted correlation between the Binder et al. (2016) conceptual norms and the norms predicted by the word embedding models. For the majority of the conceptual norms, the predictive accuracy of all three sets of word embeddings is highly similar, with little to no improvement gained from adding the OpenSubtitles and Wikipedia corpora together versus training only on either one of them. The generally high predictive value of the word embeddings for these conceptual-semantic dimensions – only for the dimensions *dark* and *slow* is the adjusted correlation for any of the sets of word embeddings lower than .6 – indicates that the word embeddings are cognitively plausible, in the sense that they characterize a semantic space that is largely consistent with human ratings of semantic dimensions. The bottom two dimensions in Figure 6.8 are not conceptual-semantic dimensions gathered from participant ratings, but word frequency measures. The decimal logarithm (log10) of word frequency is shown to be more predictable from the data, consistent with the generally accepted practice of log-transforming word frequencies when using them as predictors of behavior.

6.3.4 Effects of pseudo-conversational versus non-conversational training data on embeddings quality

The Wikipedia and OpenSubtitles corpora for the various languages included in our dataset differ in size (training corpus sizes for each language are reported online at https://github.com/jvparidon/subs2vec, where the word vectors are available for download). Because the size of the training corpus has been demonstrated to affect the quality of word embeddings (see Mandera et al., 2017, for example), it is crucial to correct for corpus size when drawing conclusions about

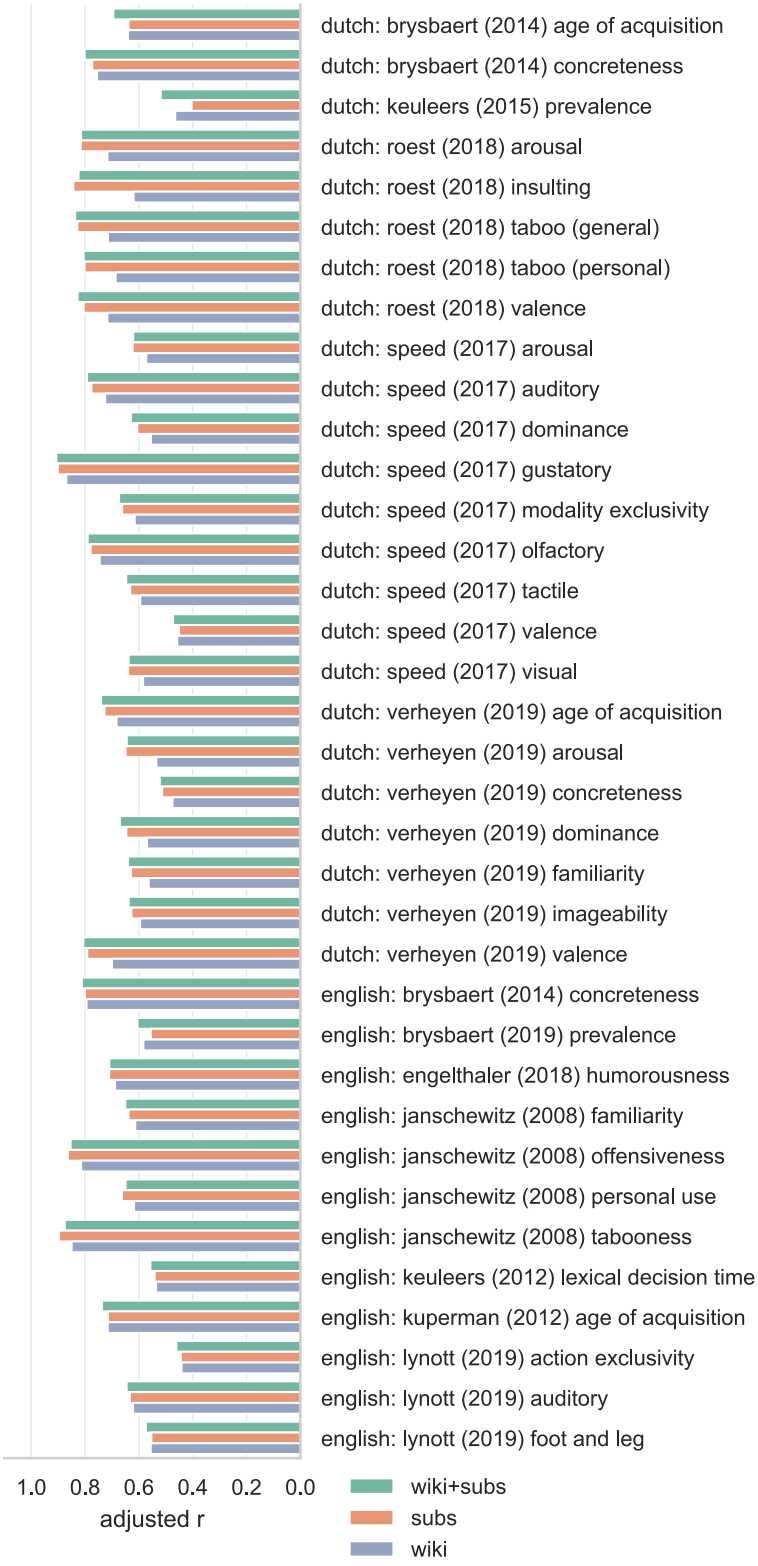


Figure 6.3: Correlations between lexical norms and our predictions for those norms based on cross-validated ridge regression using word vectors. Correlations are adjusted by penalizing for missing word vectors. 1/4

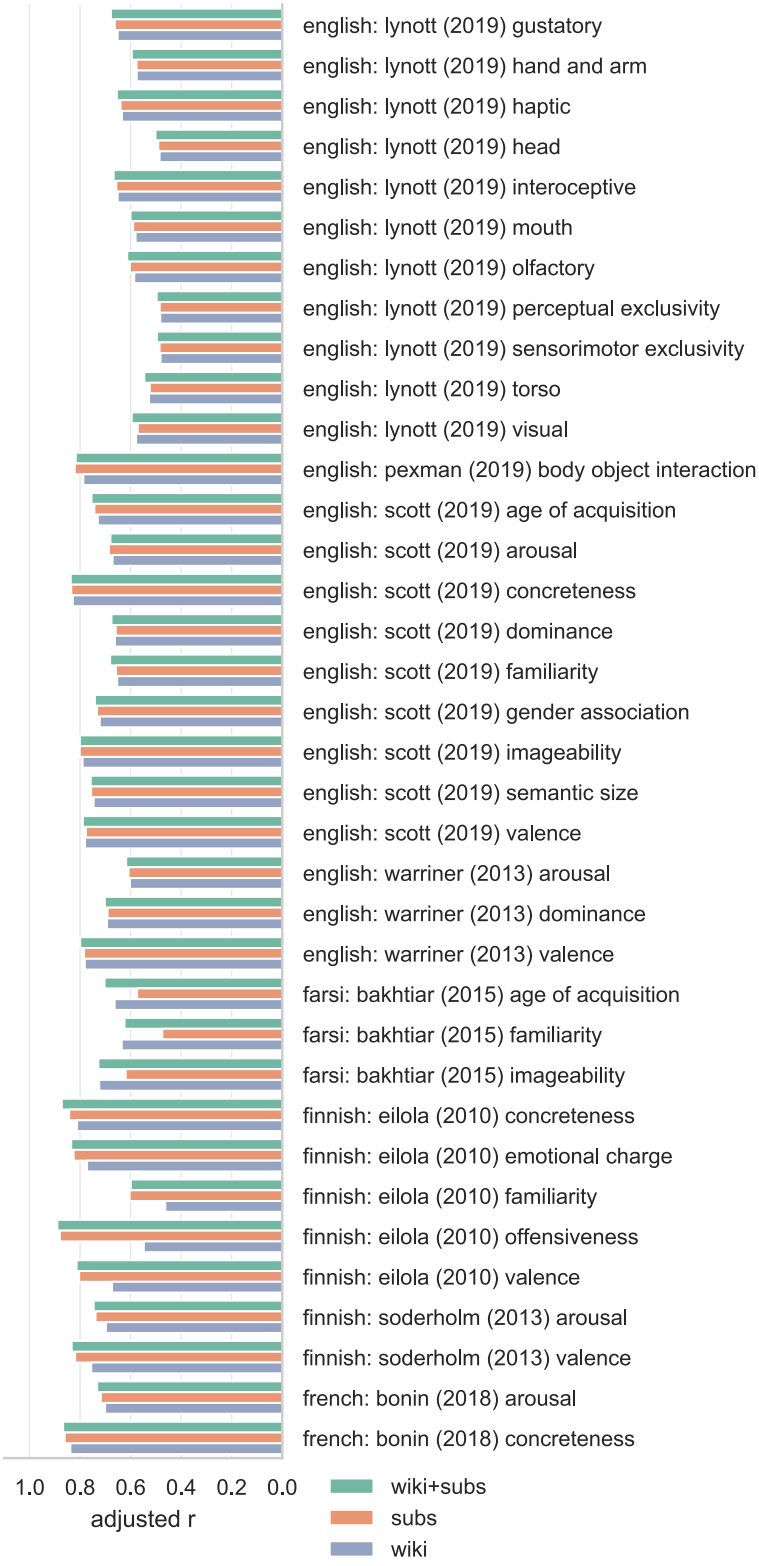


Figure 6.4: Correlations between lexical norms and our predictions for those norms based on cross-validated ridge regression using word vectors. Correlations are adjusted by penalizing for missing word vectors. 2/4

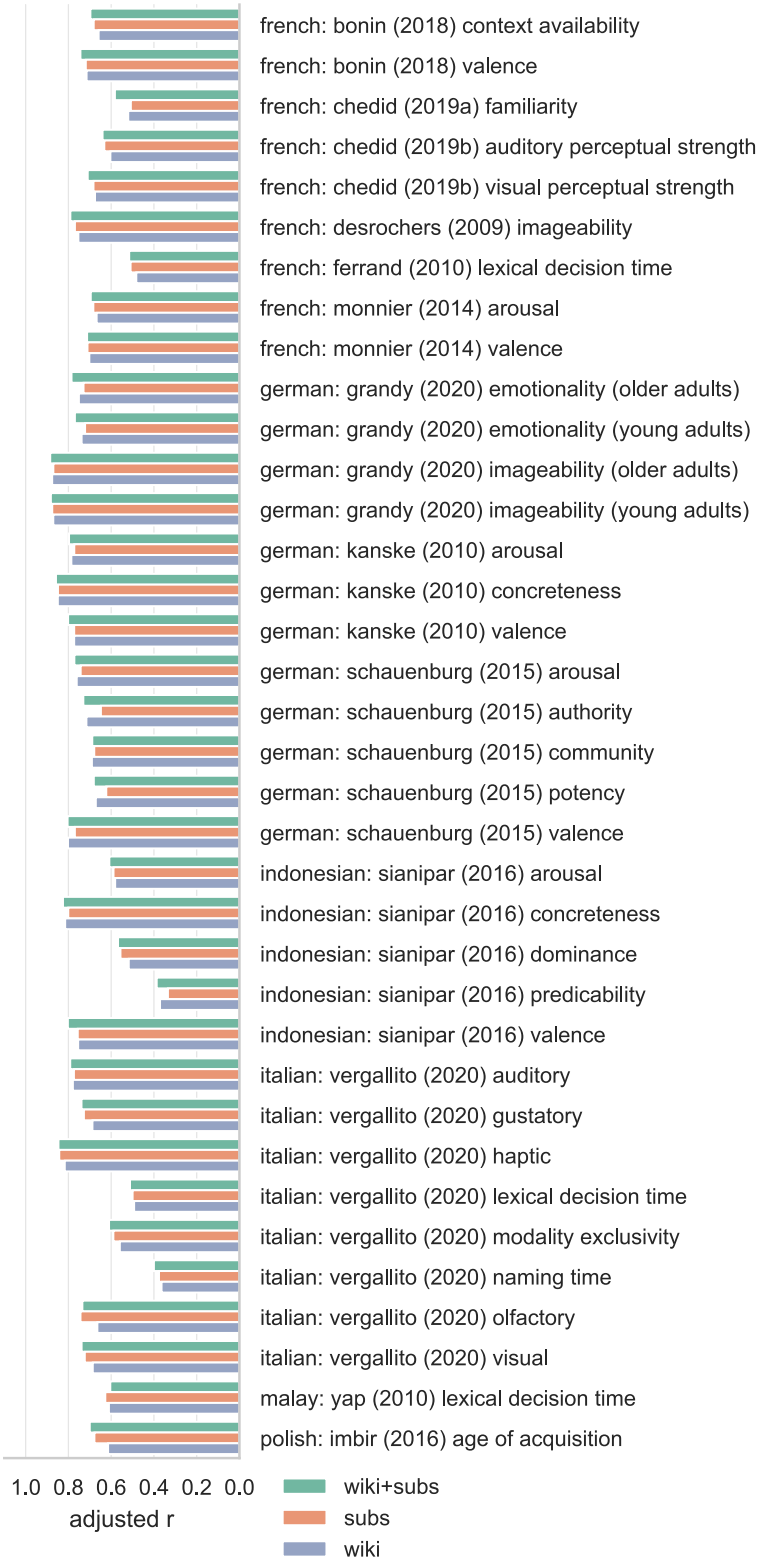


Figure 6.5: Correlations between lexical norms and our predictions for those norms based on cross-validated ridge regression using word vectors. Correlations are adjusted by penalizing for missing word vectors. 3/4

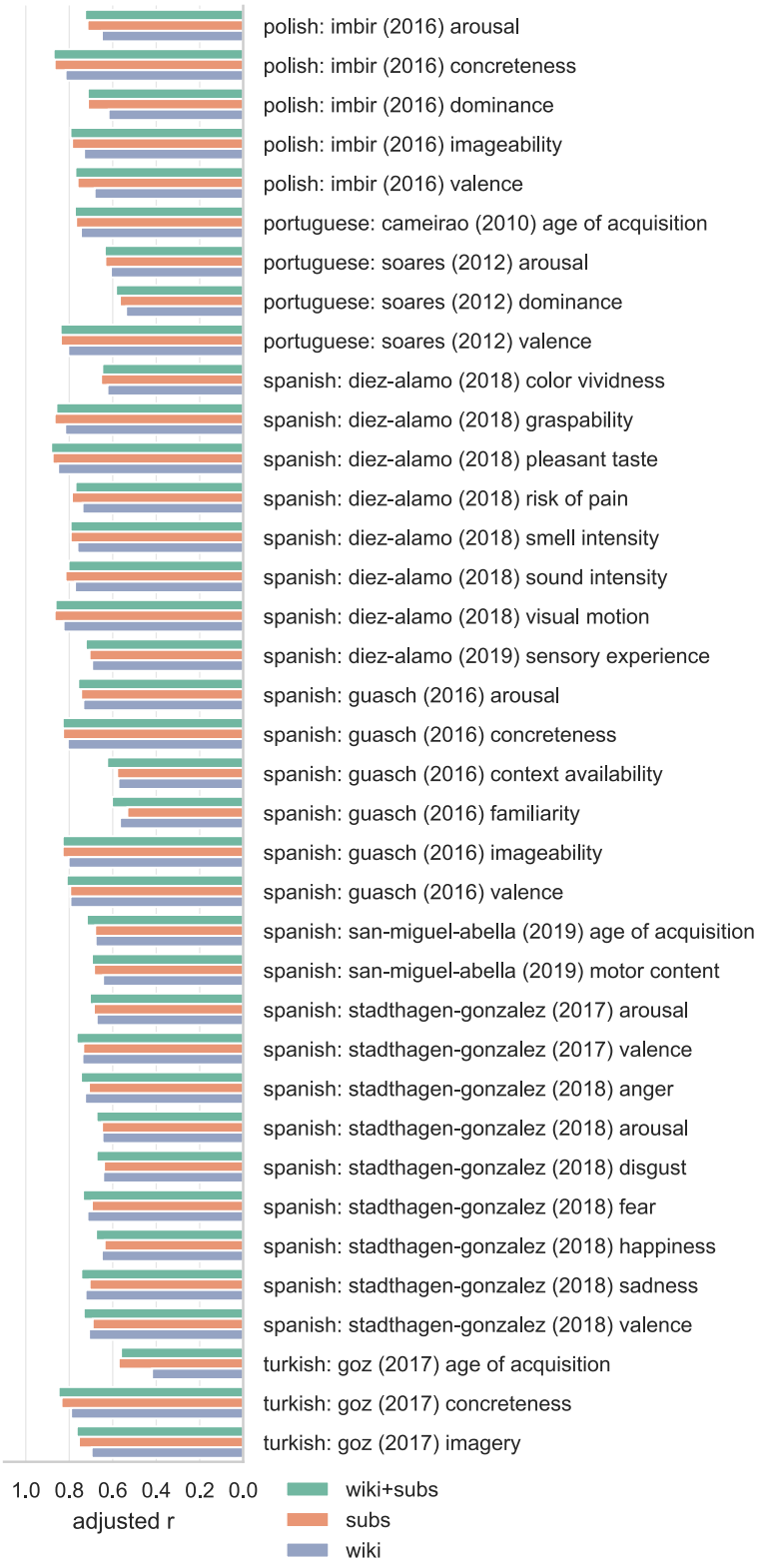


Figure 6.6: Correlations between lexical norms and our predictions for those norms based on cross-validated ridge regression using word vectors. Correlations are adjusted by penalizing for missing word vectors. 4/4

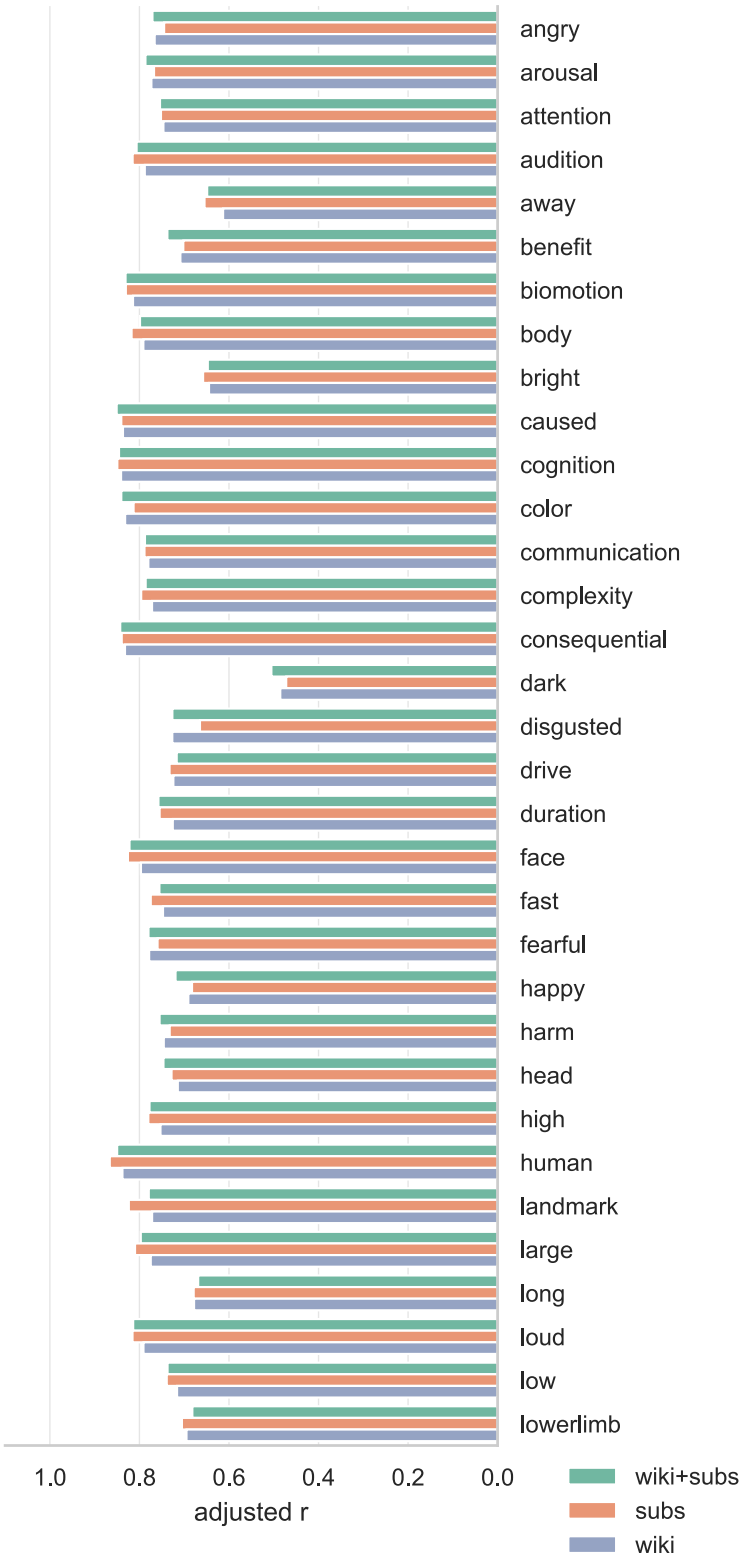


Figure 6.7: Correlations between Binder conceptual norms and our predictions for those norms based on cross-validated ridge regression using word vectors. Correlations are adjusted by penalizing for missing word vectors. 1/2

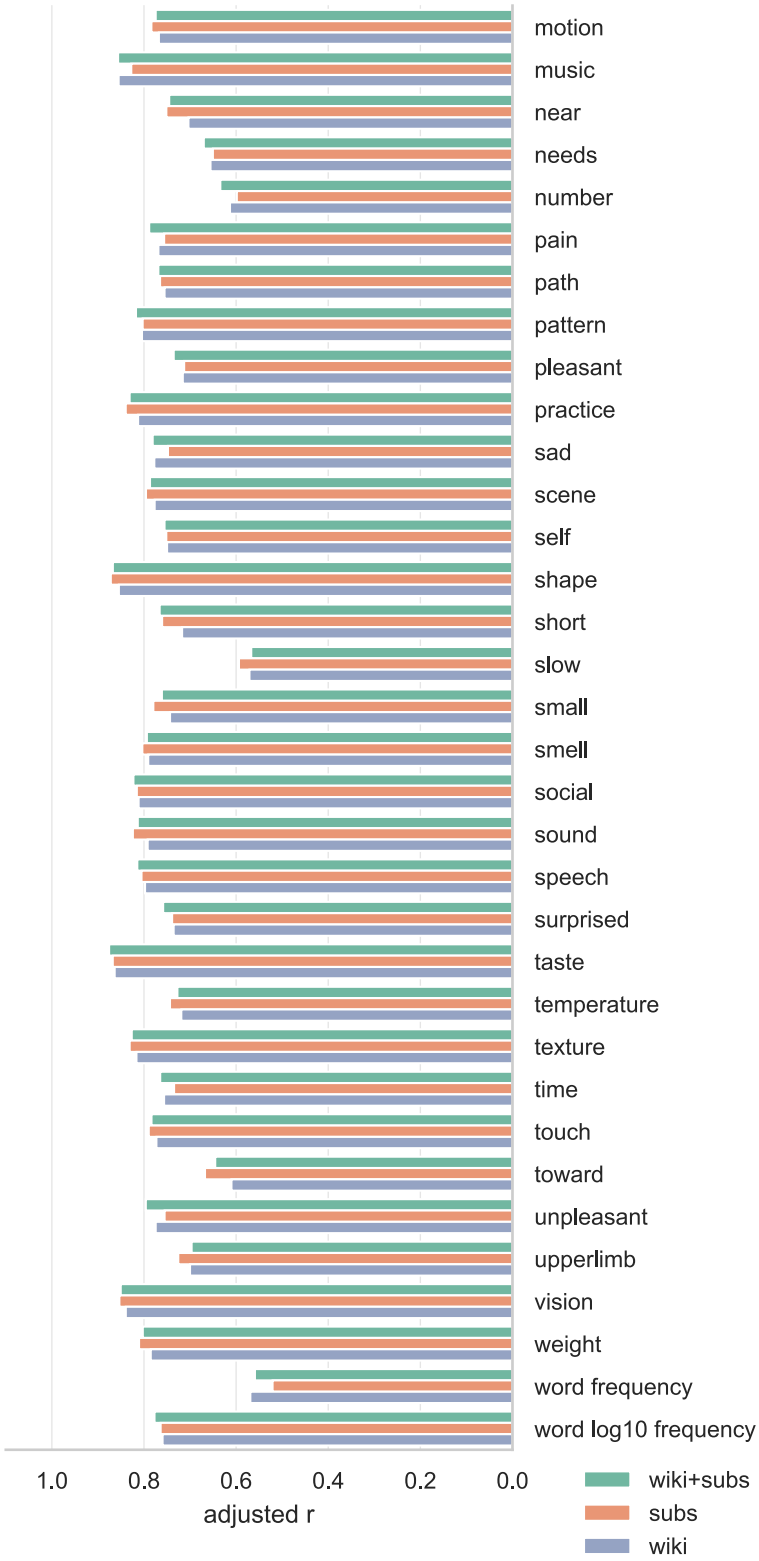


Figure 6.8: Correlations between Binder conceptual norms and our predictions for those norms based on cross-validated ridge regression using word vectors. Correlations are adjusted by penalizing for missing word vectors. 2/2

the relative merits of subtitles versus Wikipedia as training corpora. In Figure 6.9, training corpus word count-adjusted mean scores per language for each task (semantic similarities, solving analogies, and lexical norm prediction) are shown for subtitle word embeddings versus Wikipedia word embeddings. Scores were adjusted by dividing them by the log-transformed word count of their respective training corpus.

Points above the diagonal line in the figure represent relatively better performance for pseudo-conversational data, whereas points below the line represent better performance for non-conversational data. For the similarities and norms tasks the majority of points fall above the diagonal. For the analogies about half the points fall below the diagonal, but these points specifically represent the languages for which the semantic analogies dataset contain the aforementioned bias towards obscure geographic knowledge, whereas for all of the languages (Dutch, Greek, and Hebrew) for which we constructed a more psychologically plausible semantic dataset (the *no geo* datasets) the points fall above the diagonal. Overall, points fall fairly close to the diagonal, indicating that differences in performance between the subtitle and Wikipedia embeddings are relatively minor.

To test the effect of the different training corpora on embedding quality statistically we conducted a Bayesian multilevel Beta regression, with training corpus size, training corpus type, evaluation task, and the interaction of training corpus type and evaluation task as fixed effects and language and specific evaluation dataset as random intercepts. Priors on all reported coefficients were set to $\mathcal{N}(0,1)$, a mild shrinkage prior. We implemented this model in PyMC3, and sampled from it using the No-U-Turn Sampler (Hoffman & Gelman, 2014; Salvatier et al., 2016). We ran 4 chains for 2500 warmup samples each, followed by 2500 true posterior samples each (for a total of 10,000 posterior samples). Sampler diagnostics were all within acceptable limits (no divergences, $\hat{r}$ below 1.01 and at

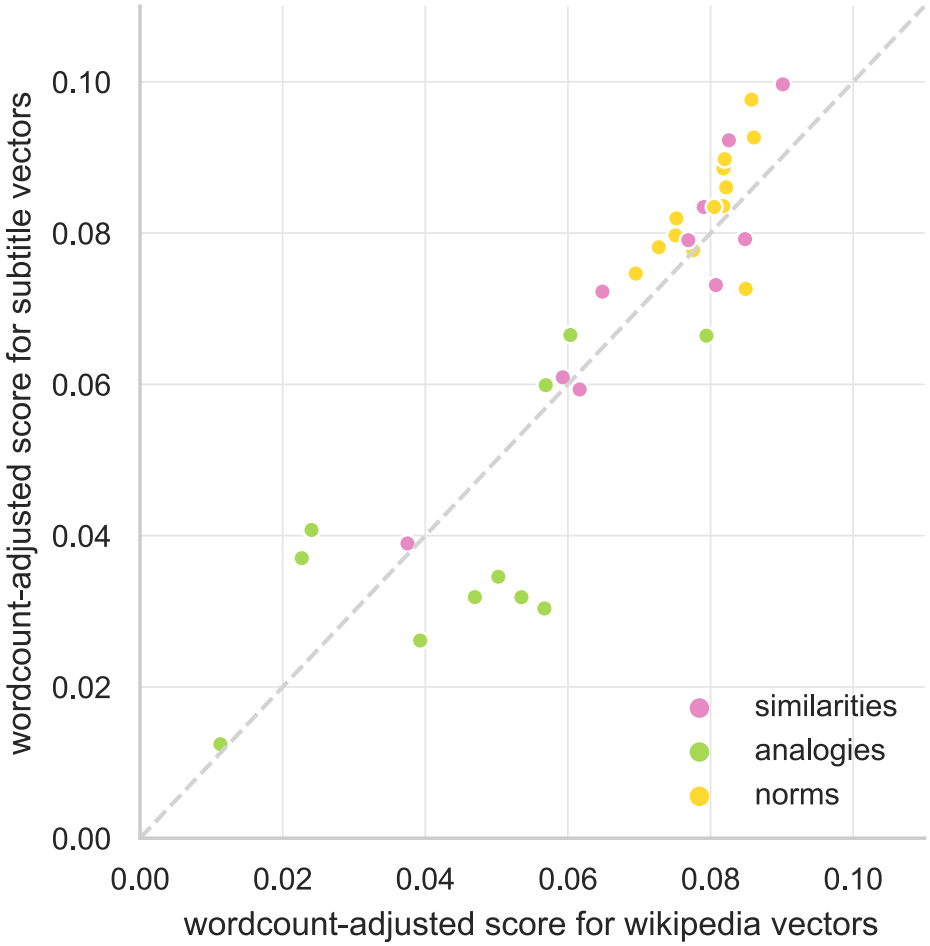


Figure 6.9: Mean evaluation scores per language and task, after correcting for training corpus size, for subtitle word embeddings versus Wikipedia word embeddings. Points above the diagonal line reflect relatively better performance for subtitle vectors than Wikipedia vectors.

least 1000 effective samples for all parameters. Further details on the inferential model, such as a directed acyclic graph of the model and trace summaries, are reported in Appendix 6.A.

This regression analysis demonstrates that after correcting for size of training corpus, subtitle embeddings are virtually indistinguishable from Wikipedia embeddings (or combined subtitle and Wikipedia embeddings) in terms of overall embedding quality (see Figure 6.10 for coefficient estimates). As is to be expected, the aforementioned advantage of a training corpus containing Wikipedia for solving geographic analogies is visible in the interaction estimates as well.

6.4 Discussion

Our aim in this study was to make available a collection of word embeddings trained on pseudo-conversational language in as many languages as possible using the same algorithm. We introduced vector embeddings in 55 languages, trained using the fastText implementation of the skipgram algorithm on the OpenSubtitles dataset. We selected the fastText algorithm because 1) it represents the state of the art in word embedding algorithms at the time of writing; and 2) there is an efficient, easy to use, and open-source implementation of the algorithm. In order to evaluate the performance of these vectors, we also trained vector embeddings on Wikipedia, and on a combination of Wikipedia and subtitles, using the same algorithm. We evaluated all of these embeddings on standard benchmark tasks. In response to the limitations of these standard evaluation tasks (Faruqui et al., 2016), we curated a dataset of multilingual lexical norms and evaluated all vector embeddings on their ability to accurately predict these ratings. We have made all of these materials, including utilities to easily obtain preprocessed versions of the original training datasets (and derived word, bigram, and trigram frequencies), available online

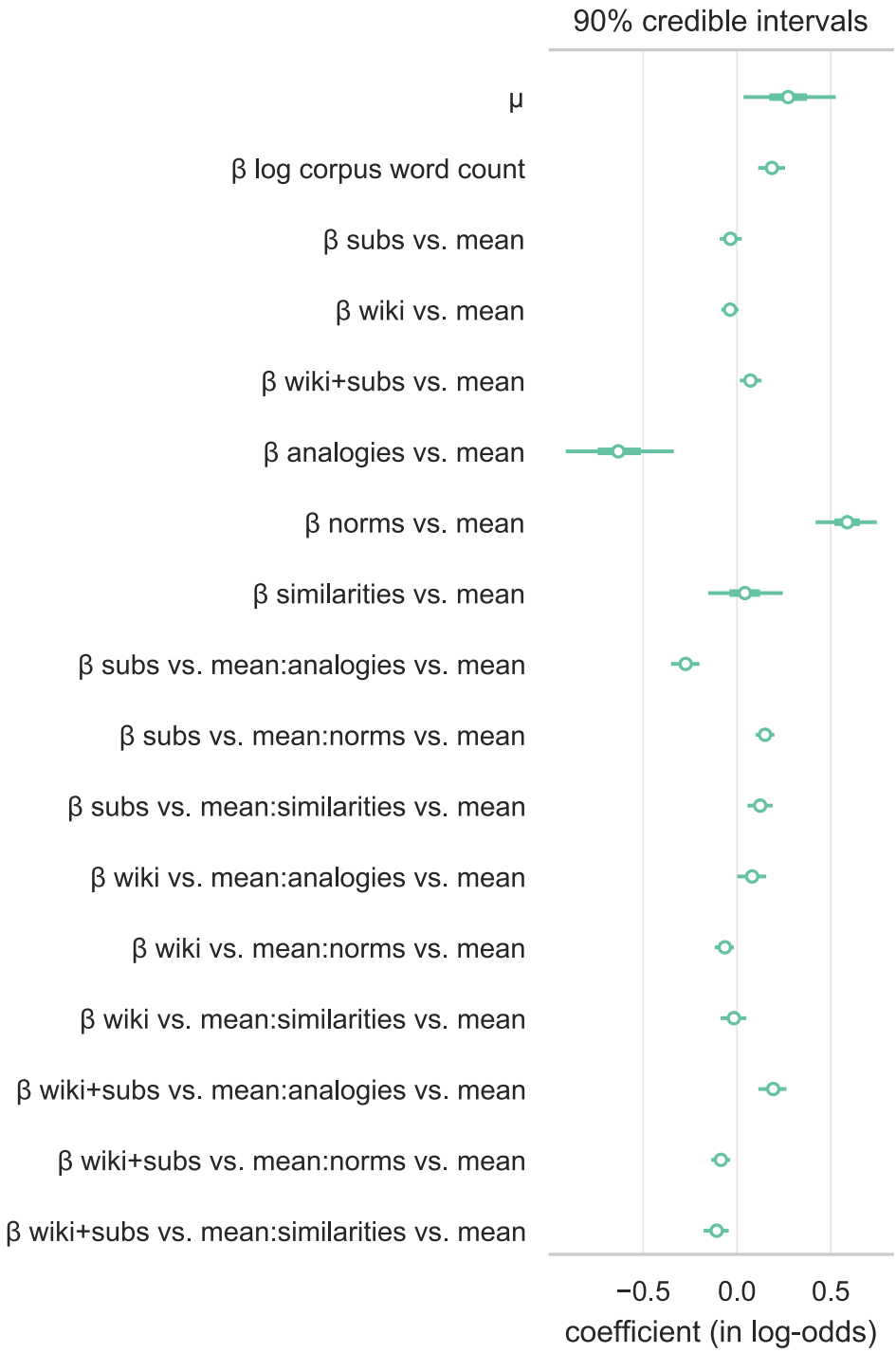


Figure 6.10: Posterior estimates from Beta regression model of OpenSubtitles and Wikipedia embeddings performance on our evaluation tasks. Beta regression uses a logit link function, therefore coefficients can be interpreted similarly to coefficients in other logit-link regressions (e.g., logistic regression). Model uses effects coding for the contrast; for example, *subs vs. mean* indicates the performance of subtitle-based embeddings relative to the mean performance of all three sets of embeddings.

at https://github.com/jvparidon/subs2vec. These materials include the full binary representations of the embeddings we trained in addition to plain-text vector representations. The binaries can be used to compute embeddings for out-of-sample vocabulary, allowing other researchers to explore the embeddings beyond the analyses reported here.

6.4.1 Performance and evaluation

Contrary to our expectations, conversational embeddings did not generally outperform alternative embeddings at predicting human lexical judgments (this contrasts with previously published predictions as well, see e.g., Mandera et al., 2017, p. 75). Our evaluation of embeddings trained on pseudo-conversational speech transcriptions (OpenSubtitles) showed that they exhibit performance rates similar to those exhibited by embeddings trained on a highly structured, knowledge-rich dataset (Wikipedia). This attests to the structured lexical relationships implicit in conversational language. However, we also suspect that more nuanced evaluation methods would reveal more substantive differences between the representations induced from these corpora. Vectors trained on pseudo-conversational text consistently outperformed vectors trained on encyclopedic text in predicting lexical judgments relating to offensiveness or tabooness, but underperformed the alternative in solving knowledge-based semantic analogies in the geographic domain (e.g., relationships between countries and capital cities). Neither of these evaluation tasks were explicitly chosen by us because they were intended to be diagnostic of one particular kind of linguistic experience, but it is notable that tabooness and offensiveness of common insults for instance are common knowledge, whereas the relationship between small countries and their respective currencies is not something the average person would know, and therefore a poor test of cognitive plausibility.

The development of evaluation tasks that are independently predicted to be solvable after exposure to conversational language merits further study.

Unfortunately, we were not able to compile evaluation metrics for every one of the 55 languages in which we provide embeddings. We did locate suitable evaluation datasets for 19 languages (and in many of these cases we provide multiple different evaluation datasets per language). That leaves embeddings in 36 languages for which we could not locate suitable evaluation datasets. This does not preclude the use of these embeddings, but we recommend researchers use them with appropriate caution, specifically by taking into account the size of the corpus that embeddings were trained on (see Appendix 6.B).

Overall, we found that embeddings trained on a combination of Wikipedia and OpenSubtitles generally outperformed embeddings trained on either of those corpora individually, even after accounting for corpus size. We hypothesize this is because the subtitle and Wikipedia embeddings represent separate, but overlapping semantic spaces, which can be jointly characterized by embeddings trained on a combined corpus. Taking into account the effect of corpus size, we recommend researchers use the embeddings trained on the largest and most diverse corpus available (subtitles plus Wikipedia, in the present study), unless they have hypotheses specific to embeddings trained on a conversational corpus.

6.4.2 Extending language coverage through complementary multilingual corpora

Our primary aim for the present study was to produce embeddings in multiple languages trained on a dataset that is more naturalistic than the widely available alternatives in multiple languages (embeddings trained on Wikipedia and other text scraped from the internet). However, it also contributes to the availability and quality of word vectors for underrepresented and less studied languages. Specifically, in some of these languages, the corresponding corpus of Wikipedia

articles is small or of low quality, while the OpenSubtitles corpus is substantially larger (e.g., Bulgarian, 4x larger; Bosnian, 7x larger; Greek, 5x larger; Croatian, 6x larger; Romanian, 7x larger; Serbian, 5x larger; Turkish, 4x larger). As a result, our study helps to increase the number of languages for which high quality embeddings are available, regardless of whether the pseudo-conversational nature of the training corpus is germane to the specific purpose for which the embeddings may be used.

6.4.3 Translation vs. original language

An important caveat in using the OpenSubtitles corpus in the present context is that many of the subtitles are translations, meaning the subtitles are not straight transcriptions, but a translation from speech in the original language a movie or television series was released in to text in another language. Moreover, while it is highly likely that translators try to produce subtitles that are correct and coherent in the target language, we have no reliable way of ascertaining the proficiency of the (often anonymous) translator in either source or language. In the present context it was not feasible to examine which parts of the subtitle corpus are translations and which represent straight transcriptions of audio in the original language and therefore we could not test whether training on translated subtitles has an adverse effect on word embedding quality. This issue is not unsolvable in principle, because the original language of the movies and television series for which each set of subtitles was written can be established using secondary, publicly available datasets. Future work investigating distributional differences between transcribed and translated dialogue seems warranted.

A related ambiguity is whether subtitles should be viewed as representing experience of written or spoken language. On the one hand, subtitles are read by many people. However, as transcriptions of speech, subtitles convey a more direct representation of spoken language experience than is conveyed by other

written corpora such as Wikipedia. This second interpretation was an important part of our motivation, but the interpretation of subtitles as written language is also important.

6.4.4 Advances in fastText algorithms

The most recent implementation of the fastText algorithm includes CBOW with position-dependent weighting of the context vectors, which seems to represent another step forward in terms of the validity of the word embeddings it generates (Mikolov et al., 2017). As of the time of writing, this implementation has not been released to the public (although a rudimentary description of the algorithm has been published, alongside a number of word vector datasets in various languages created using the new version of the algorithm). Because all the code used in the present study is publicly available, if and when an implementation of the new algorithm is released to the public, the present study and dataset can easily be reproduced using this improved method for computing word vectors.

Algorithmic developments in the field of distributional semantics move quickly. Nonetheless, in this paper we have produced (for a large set of languages, using state of the art methods) word embeddings trained on a large corpus of language that reflects real-world linguistic experience. In addition to insights about language and cognition that can be gleaned from these embeddings directly, they are a valuable resource for improving statistical models of other psychological and linguistic phenomena.

6.4.5 Open practices statement

All of the datasets and code presented in this paper, as well as the datasets and code necessary to reproduce the analyses, are freely available online at https://github.com/jvparidon/subs2vec.

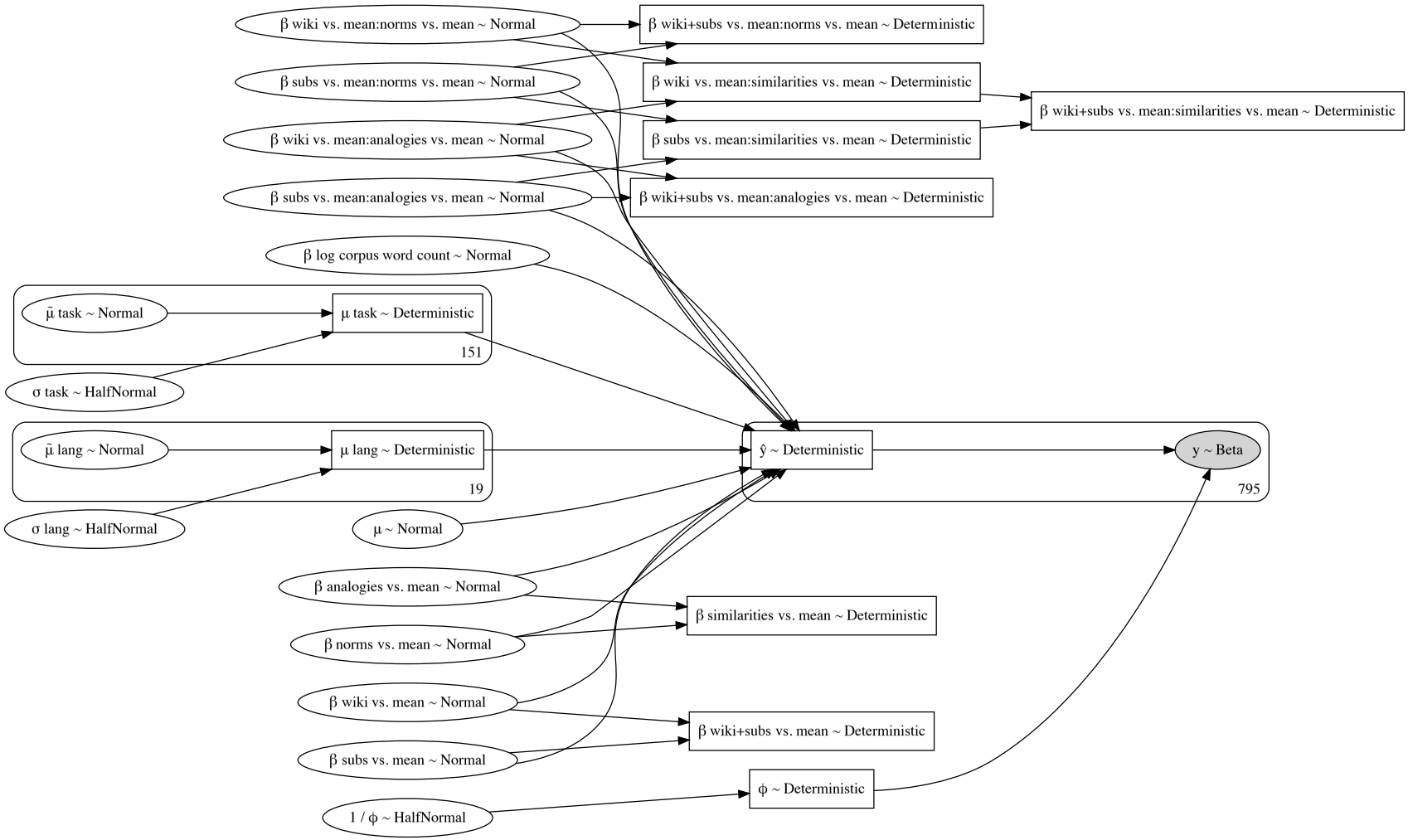
The *subs2vec* Python package also provides tools can be used to compute semantic dissimilarities, solve analogies, and predict lexical norms for novel datasets.

Appendix 6.A Inferential model details

Table 6.5: Summary of posterior traces for inferential model. 90% CI upper and lower refer to upper and lower bounds of the credible interval, n_{eff} is the estimated effective sample size.

	mean	sd	90% CI lower	90% CI upper	n_{eff} mean	n_{eff} sd	$\hat{r}$
μ	0.28	0.15	0.04	0.51	1382.0	1382.0	1.00
β log corpus word count	0.19	0.04	0.12	0.26	4159.0	4159.0	1.00
β wiki vs. mean	-0.04	0.03	-0.08	0.01	4716.0	4716.0	1.00
β subs vs. mean	-0.04	0.04	-0.09	0.02	4237.0	4237.0	1.00
β norms vs. mean	0.59	0.10	0.43	0.76	1680.0	1672.0	1.00
β analogies vs. mean	-0.62	0.17	-0.90	-0.34	2200.0	2200.0	1.00
β wiki vs. mean:norms vs. mean	-0.06	0.03	-0.12	-0.02	4425.0	4425.0	1.00
β wiki vs. mean:analogies vs. mean	0.08	0.05	0.01	0.16	5729.0	5729.0	1.00
β subs vs. mean:norms vs. mean	0.15	0.03	0.10	0.20	4977.0	4977.0	1.00
β subs vs. mean:analogies vs. mean	-0.27	0.05	-0.35	-0.20	6110.0	6104.0	1.00
β wiki+subs vs. mean	0.07	0.04	0.01	0.13	6077.0	5947.0	1.00
β similarities vs. mean	0.04	0.12	-0.16	0.24	1986.0	1986.0	1.00
β wiki+subs vs. mean:norms vs. mean	-0.08	0.03	-0.14	-0.04	12024.0	11067.0	1.00
β wiki+subs vs. mean:analogies vs. mean	0.19	0.05	0.11	0.26	14384.0	13239.0	1.00
β subs vs. mean:similarities vs. mean	0.12	0.04	0.05	0.19	13000.0	11975.0	1.00
β wiki vs. mean:similarities vs. mean	-0.02	0.04	-0.09	0.05	13382.0	6400.0	1.00
β wiki+subs vs. mean:similarities vs. mean	-0.11	0.04	-0.18	-0.04	18032.0	14781.0	1.00
σ task	0.52	0.04	0.46	0.58	1907.0	1907.0	1.00
σ lang	0.45	0.10	0.28	0.61	2503.0	2503.0	1.00
ϕ	34.65	1.96	31.44	37.87	6631.0	6601.0	1.00

Figure 6.11: Directed acyclic graph of inferential model, node labels include shape of prior distribution. Random intercepts were estimated by language, but also by evaluation task where appropriate (e.g., the MC30 similarities were used in Arabic, Dutch, English, Romanian, and Spanish). The likelihood uses the Beta(μ, ϕ) parametrization of the Beta distribution. Coefficients labeled "Deterministic" follow trivially from the other coefficient estimates and were computed during model estimation.



Appendix 6.B Training corpus details

Table 6.6: Descriptive statistics for training corpora.

language	corpus	word count	mean words per line
Afrikaans	OpenSubtitles	324K	6.61
	Wikipedia	17M	17.01
	Wikipedia + OpenSubtitles	17M	16.53
Albanian	OpenSubtitles	12M	6.65
	Wikipedia	18M	16.90
	Wikipedia + OpenSubtitles	30M	10.47
Arabic	OpenSubtitles	188M	5.64
	Wikipedia	120M	18.32
	Wikipedia + OpenSubtitles	308M	7.72
Armenian	OpenSubtitles	24K	6.06
	Wikipedia	38M	21.66
	Wikipedia + OpenSubtitles	39M	21.62
Basque	OpenSubtitles	3M	4.97
	Wikipedia	20M	11.39
	Wikipedia + OpenSubtitles	24M	9.60
Bengali	OpenSubtitles	2M	5.39
	Wikipedia	19M	27.64
	Wikipedia + OpenSubtitles	21M	19.16
Bosnian	OpenSubtitles	92M	6.34
	Wikipedia	13M	13.15
	Wikipedia + OpenSubtitles	105M	6.78
Breton	OpenSubtitles	111K	5.97
	Wikipedia	8M	15.72
	Wikipedia + OpenSubtitles	8M	15.36
Bulgarian	OpenSubtitles	247M	6.87
	Wikipedia	53M	15.82
	Wikipedia + OpenSubtitles	300M	7.64
Catalan	OpenSubtitles	3M	6.95
	Wikipedia	176M	20.75
	Wikipedia + OpenSubtitles	179M	20.06
Croatian	OpenSubtitles	242M	6.44
	Wikipedia	43M	12.25
	Wikipedia + OpenSubtitles	285M	6.94

Czech	OpenSubtitles	249M	6.43
	Wikipedia	100M	13.44
	Wikipedia + OpenSubtitles	349M	7.57
Danish	OpenSubtitles	87M	6.96
	Wikipedia	56M	14.72
	Wikipedia + OpenSubtitles	143M	8.77
Dutch	OpenSubtitles	265M	7.39
	Wikipedia	249M	14.40
	Wikipedia + OpenSubtitles	514M	9.67
English	OpenSubtitles	751M	8.22
	Wikipedia	2B	17.57
	Wikipedia + OpenSubtitles	3B	13.90
Esperanto	OpenSubtitles	382K	5.44
	Wikipedia	38M	14.64
	Wikipedia + OpenSubtitles	38M	14.39
Estonian	OpenSubtitles	60M	5.99
	Wikipedia	29M	10.38
	Wikipedia + OpenSubtitles	90M	6.94
Farsi	OpenSubtitles	45M	6.39
	Wikipedia	87M	17.36
	Wikipedia + OpenSubtitles	132M	10.92
Finnish	OpenSubtitles	117M	5.10
	Wikipedia	74M	10.80
	Wikipedia + OpenSubtitles	191M	6.40
French	OpenSubtitles	336M	8.31
	Wikipedia	724M	19.54
	Wikipedia + OpenSubtitles	1B	13.69
Galician	OpenSubtitles	2M	6.58
	Wikipedia	40M	18.56
	Wikipedia + OpenSubtitles	42M	17.30
Georgian	OpenSubtitles	1M	5.21
	Wikipedia	15M	11.04
	Wikipedia + OpenSubtitles	16M	10.26
German	OpenSubtitles	139M	7.01
	Wikipedia	976M	14.06
	Wikipedia + OpenSubtitles	1B	12.49
Greek	OpenSubtitles	271M	6.90
	Wikipedia	58M	18.26
	Wikipedia + OpenSubtitles	329M	7.76
Hebrew	OpenSubtitles	170M	6.22
	Wikipedia	133M	13.92

	Wikipedia + OpenSubtitles	303M	8.22
Hindi	OpenSubtitles	660K	6.77
	Wikipedia	31M	33.89
	Wikipedia + OpenSubtitles	32M	31.28
Hungarian	OpenSubtitles	228M	6.04
	Wikipedia	121M	12.37
	Wikipedia + OpenSubtitles	349M	7.34
Icelandic	OpenSubtitles	7M	6.08
	Wikipedia	7M	13.17
	Wikipedia + OpenSubtitles	15M	8.26
Indonesian	OpenSubtitles	65M	6.18
	Wikipedia	69M	14.09
	Wikipedia + OpenSubtitles	134M	8.70
Italian	OpenSubtitles	278M	7.43
	Wikipedia	476M	18.87
	Wikipedia + OpenSubtitles	754M	12.05
Kazakh	OpenSubtitles	13K	3.90
	Wikipedia	18M	10.39
	Wikipedia + OpenSubtitles	18M	10.38
Korean	OpenSubtitles	7M	4.30
	Wikipedia	63M	11.97
	Wikipedia + OpenSubtitles	70M	10.19
Latvian	OpenSubtitles	2M	5.10
	Wikipedia	14M	10.91
	Wikipedia + OpenSubtitles	16M	9.46
Lithuanian	OpenSubtitles	6M	4.89
	Wikipedia	23M	11.10
	Wikipedia + OpenSubtitles	29M	8.74
Macedonian	OpenSubtitles	20M	6.33
	Wikipedia	27M	16.82
	Wikipedia + OpenSubtitles	47M	9.82
Malay	OpenSubtitles	12M	5.88
	Wikipedia	29M	14.50
	Wikipedia + OpenSubtitles	41M	10.11
Malayalam	OpenSubtitles	2M	4.08
	Wikipedia	10M	9.18
	Wikipedia + OpenSubtitles	12M	7.92
Norwegian	OpenSubtitles	46M	6.69
	Wikipedia	91M	14.53
	Wikipedia + OpenSubtitles	136M	10.44

Polish	OpenSubtitles	250M	6.15
	Wikipedia	232M	12.63
	Wikipedia + OpenSubtitles	483M	8.17
Portuguese	OpenSubtitles	258M	7.40
	Wikipedia	238M	18.60
	Wikipedia + OpenSubtitles	496M	10.41
Romanian	OpenSubtitles	435M	7.70
	Wikipedia	65M	16.16
	Wikipedia + OpenSubtitles	500M	8.27
Russian	OpenSubtitles	152M	6.43
	Wikipedia	391M	13.96
	Wikipedia + OpenSubtitles	543M	10.51
Serbian	OpenSubtitles	344M	6.57
	Wikipedia	70M	12.97
	Wikipedia + OpenSubtitles	413M	7.16
Sinhala	OpenSubtitles	3M	5.34
	Wikipedia	6M	14.52
	Wikipedia + OpenSubtitles	9M	8.89
Slovak	OpenSubtitles	47M	6.23
	Wikipedia	29M	12.85
	Wikipedia + OpenSubtitles	76M	7.73
Slovenian	OpenSubtitles	107M	6.15
	Wikipedia	32M	13.45
	Wikipedia + OpenSubtitles	138M	7.02
Spanish	OpenSubtitles	514M	7.46
	Wikipedia	586M	20.36
	Wikipedia + OpenSubtitles	1B	11.25
Swedish	OpenSubtitles	101M	6.87
	Wikipedia	143M	11.93
	Wikipedia + OpenSubtitles	245M	9.15
Tagalog	OpenSubtitles	88K	6.02
	Wikipedia	7M	17.16
	Wikipedia + OpenSubtitles	7M	16.74
Tamil	OpenSubtitles	123K	4.36
	Wikipedia	17M	10.09
	Wikipedia + OpenSubtitles	17M	10.00
Telugu	OpenSubtitles	103K	4.50
	Wikipedia	15M	10.34
	Wikipedia + OpenSubtitles	15M	10.25
Turkish	OpenSubtitles	240M	5.56
	Wikipedia	55M	12.52

	Wikipedia + OpenSubtitles	295M	6.20
Ukrainian	OpenSubtitles	5M	5.51
	Wikipedia	163M	13.34
	Wikipedia + OpenSubtitles	168M	12.80
Urdu	OpenSubtitles	196K	7.02
	Wikipedia	16M	28.88
	Wikipedia + OpenSubtitles	16M	27.83
Vietnamese	OpenSubtitles	27M	8.23
	Wikipedia	115M	20.51
	Wikipedia + OpenSubtitles	143M	15.94

Appendix 6.C Unpenalized evaluation scores

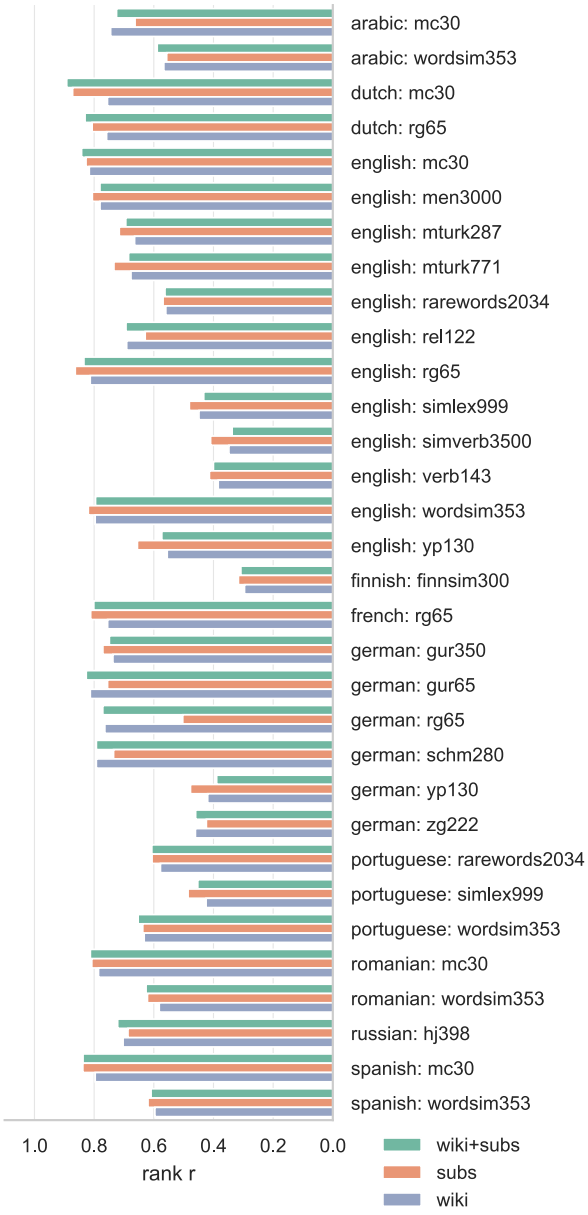


Figure 6.12: Unpenalized rank correlations between human ratings of semantic similarity and word vector cosine similarity.

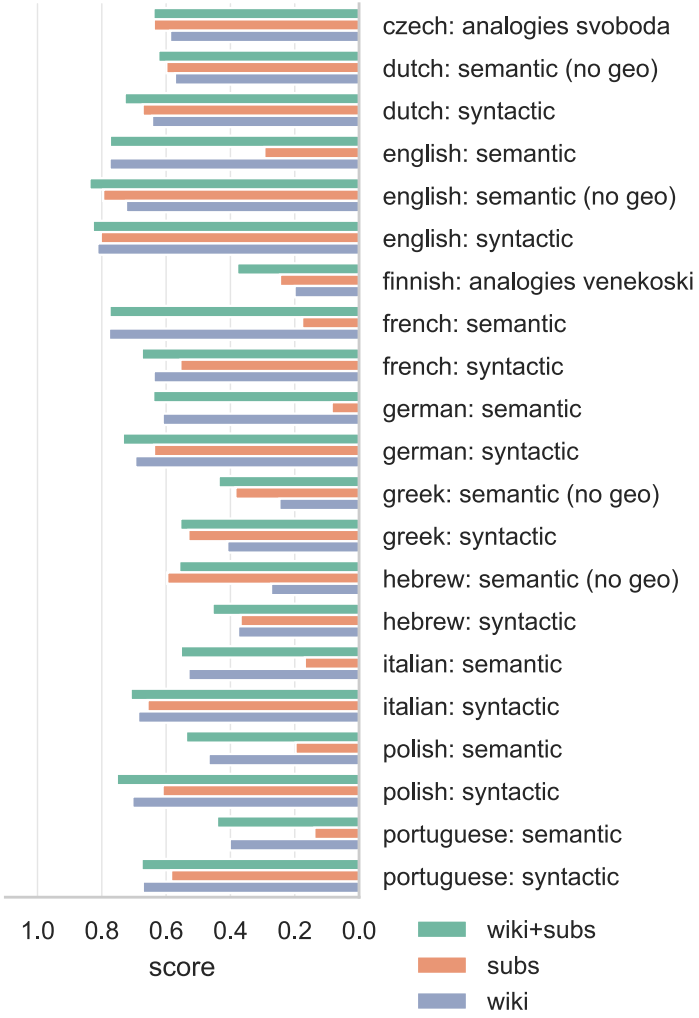


Figure 6.13: Unpenalized proportion of correctly solved analogies in the semantic and syntactic domain using word vectors. Semantic datasets contained 93% geographic analogies, *no geo* datasets are those same datasets, excluding the geographic analogies.

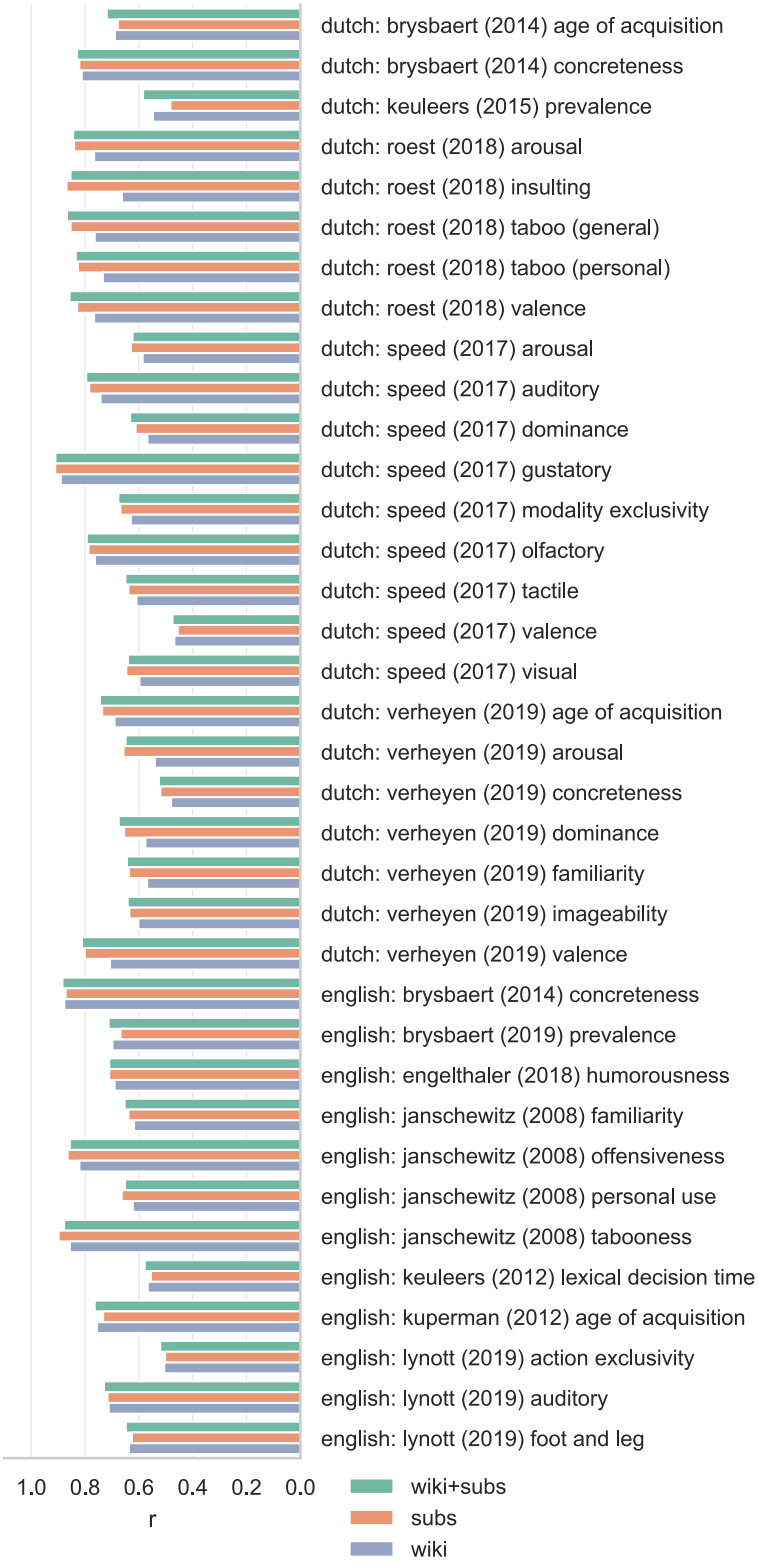


Figure 6.14: Unpenalized correlations between lexical norms and our predictions for those norms based on cross-validated ridge regression using word vectors. 1/4

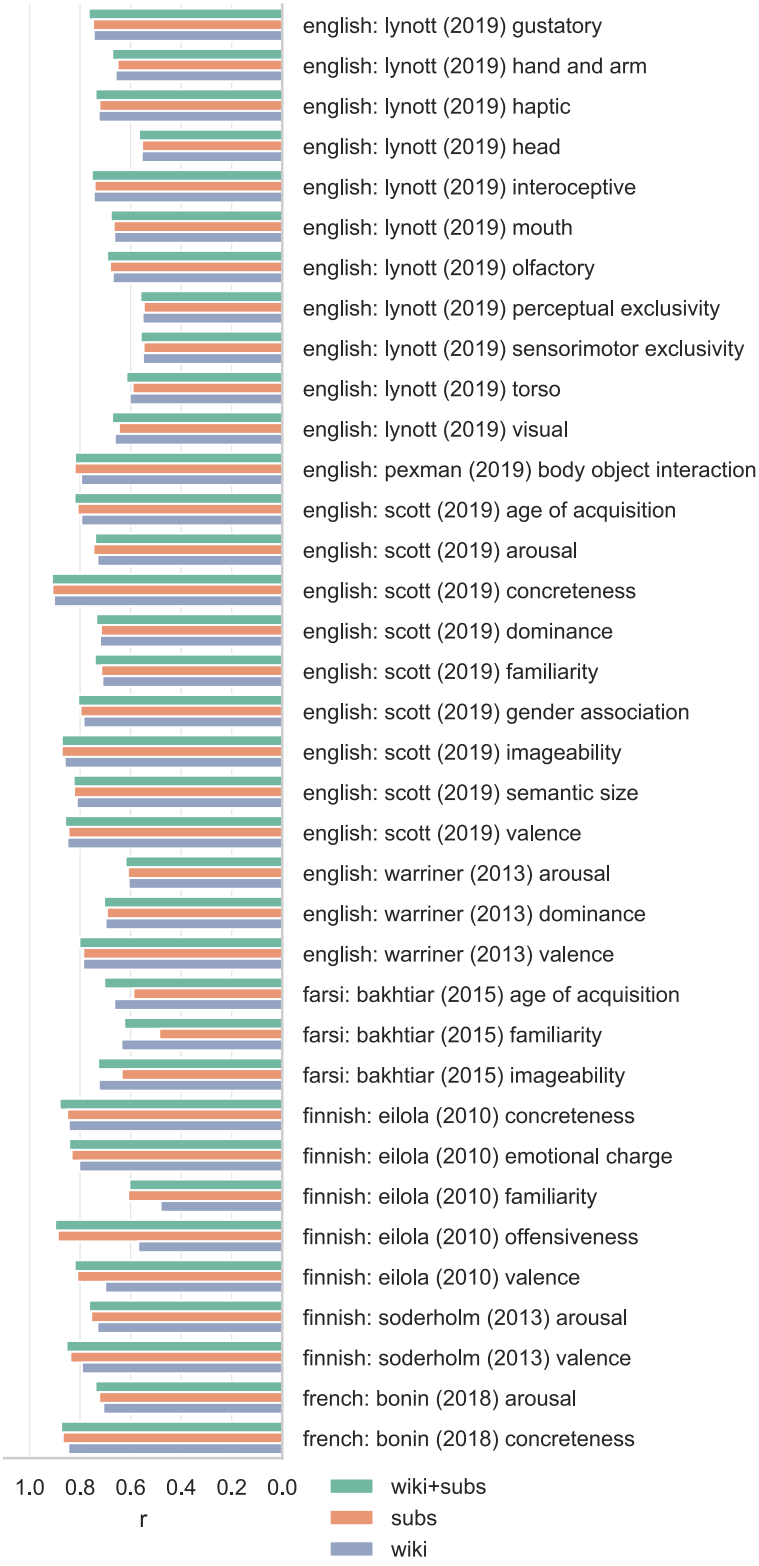


Figure 6.15: Unpenalized correlations between lexical norms and our predictions for those norms based on cross-validated ridge regression using word vectors. 2/4

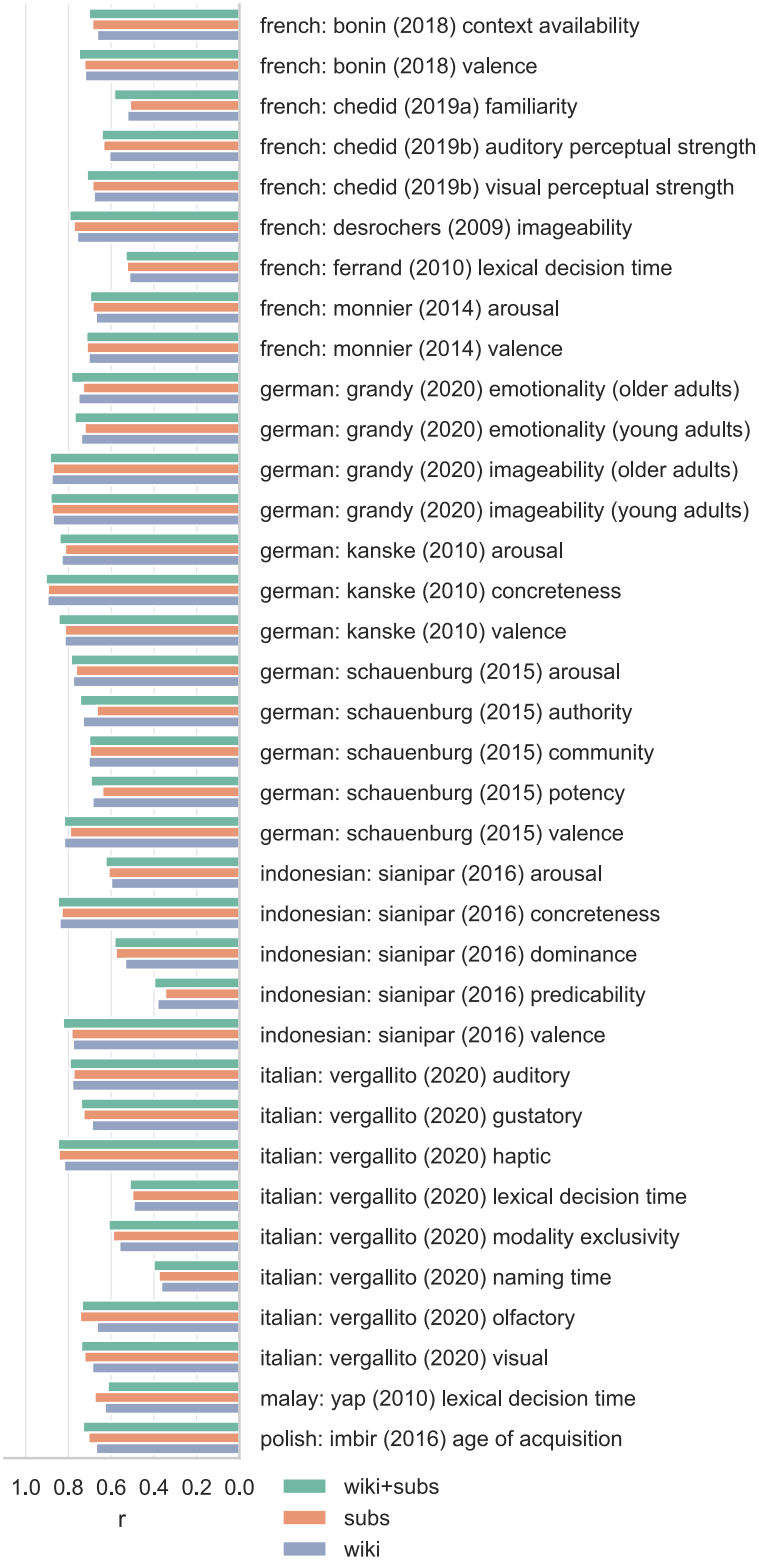


Figure 6.16: Unpenalized correlations between lexical norms and our predictions for those norms based on cross-validated ridge regression using word vectors. 3/4

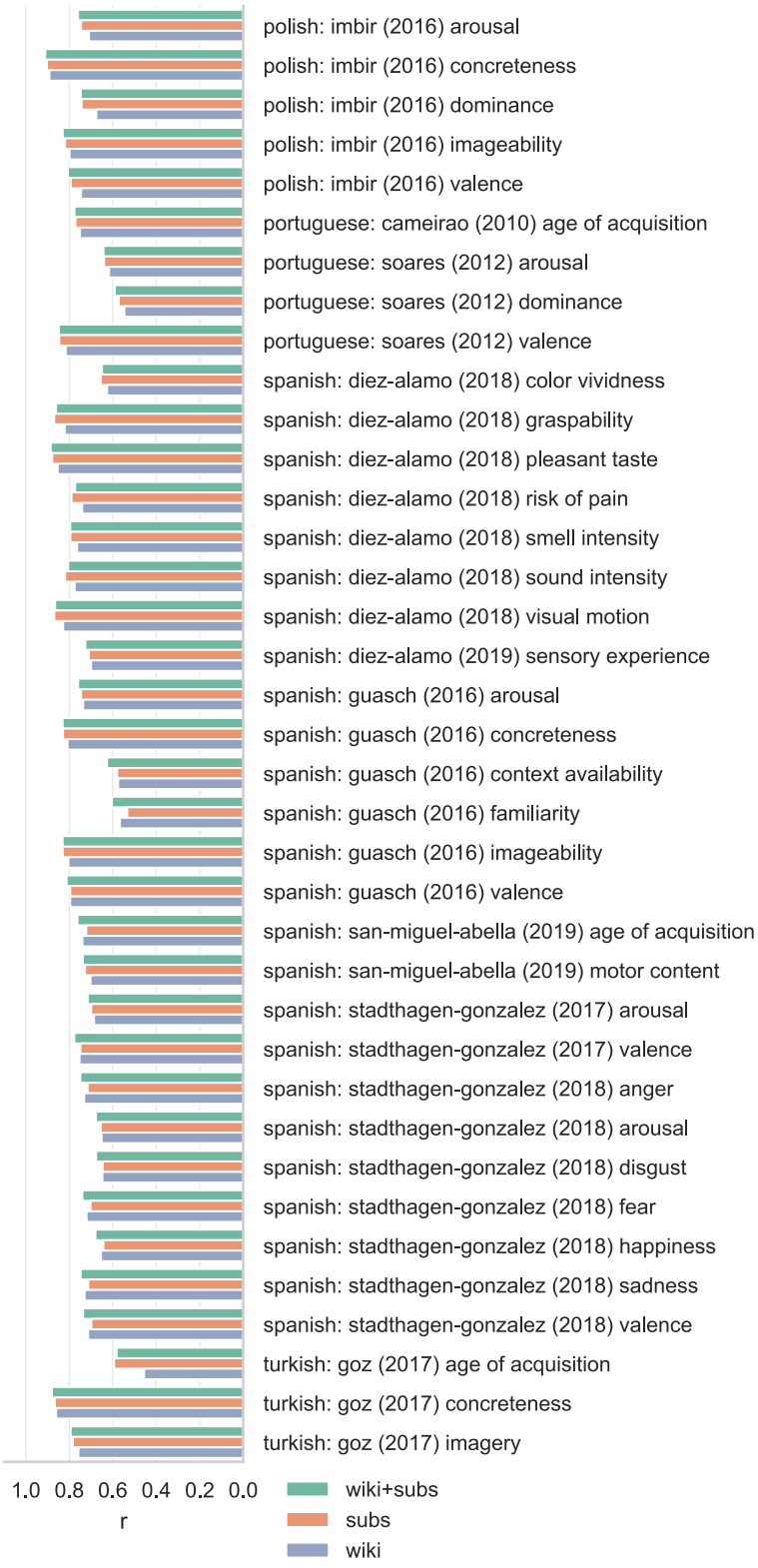


Figure 6.17: Unpenalized correlations between lexical norms and our predictions for those norms based on cross-validated ridge regression using word vectors. 4/4

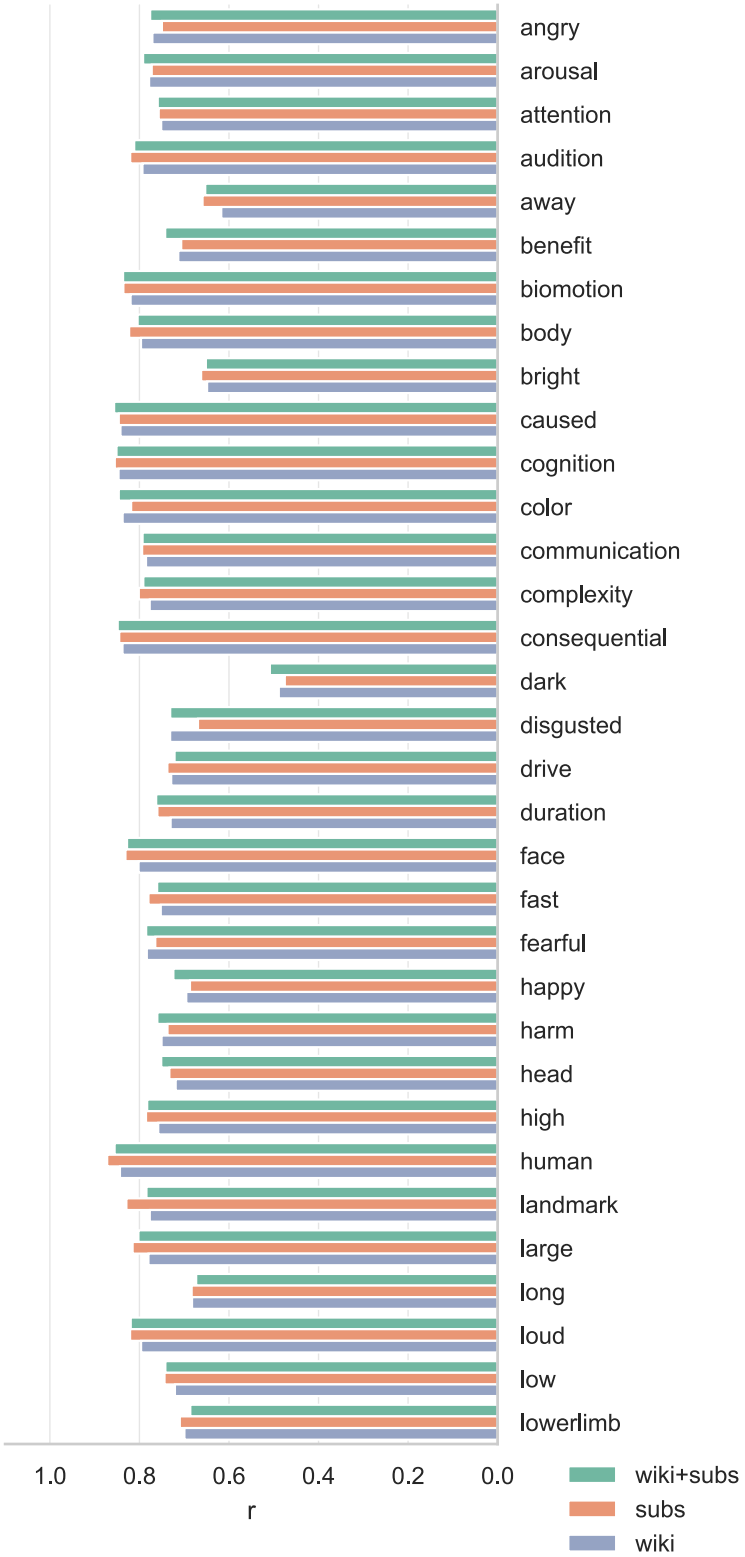


Figure 6.18: Unpenalized correlations between Binder conceptual norms and our predictions for those norms based on cross-validated ridge regression using word vectors. 1/2

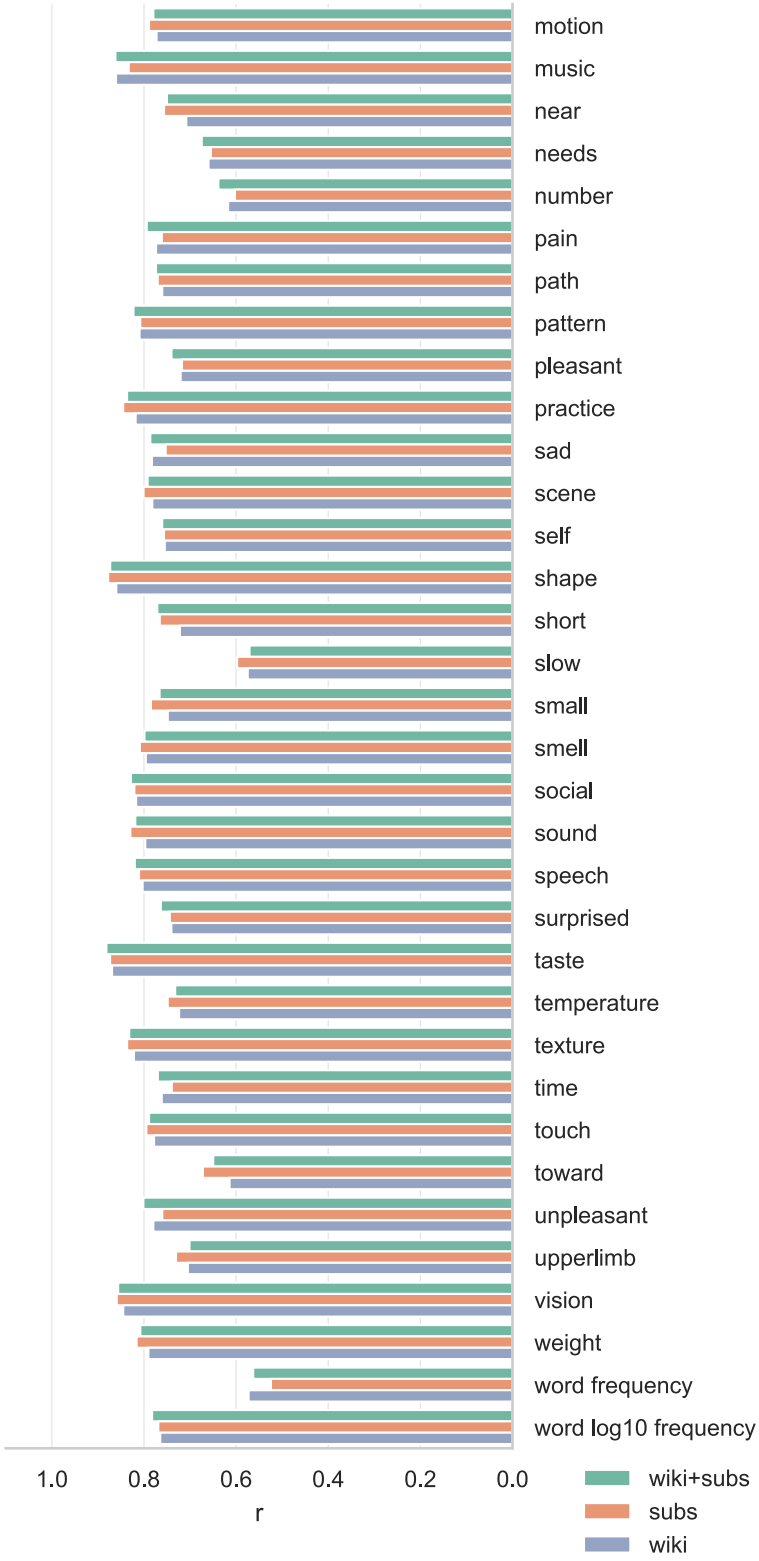


Figure 6.19: Unpenalized correlations between Binder conceptual norms and our predictions for those norms based on cross-validated ridge regression using word vectors. 2/2

7 | Summary and discussion

As noted in Chapter 1, in one of the very first experimental studies of concurrent speech comprehension and production, Broadbent (1952) posited that in order to perform concurrent comprehension and production (as in simultaneous interpreting) comprehension and production processes need to be independent enough not to interfere with each other, yet a large body of psycholinguistic literature tells us comprehension and production rely on shared resources and mechanisms. In this dissertation, I set out to provide an account of temporal coordination of comprehension and production processes in simultaneous interpreting and speech shadowing that integrates prior literature on speech production and comprehension with data gathered in behavioral studies of simultaneous interpreting. This account was to be computationally implemented so as to provide falsifiable quantitative predictions, giving it explanatory power beyond the existing box-and-arrow process models of comprehension and production in simultaneous interpreting.

7.1 Chapter summaries

In Chapter 2 I described the proposed model of interpreting and shadowing: a process model based on the process hierarchy and process durations from Indefrey and Levelt's (2004) meta-analysis of word production and comprehension studies. This model was computationally implemented as a series of simplified linear ballistic accumulators (Brown & Heathcote, 2008), essentially turning each

box in the box-and-arrow model from Indefrey and Levelt (2004) into a fixed-duration stage that a word has to pass through before it can continue to the next stage. I conducted a behavioral experiment using untrained participants, asking them to shadow and interpret simple prerecorded narratives at speech rates ranging from 100 words per minute to 200 words per minute, and compared the observed error rates to those predicted by the computational model. Only when I introduced modifications to the model in the form of an additional, low-level shadowing route and a switch cost for switching access to lexical selection between comprehension and production, could I obtain a good fit between the computational model's predictions to the observed error rates.

Both of these modifications are compatible with prior findings in the literature: Evidence for a low-level route for shadowing can be found in Marslen-Wilson's (1973, 1975, 1985) seminal work on close speech shadowing, although he also emphasized that even at very low input-output latencies, participants can still recall some content and sometimes make simple repairs to errors in the input speech signal, implying that at least some of the signal reaches processing stages beyond the low-level shadowing route I propose in the computational model. Switch costs are well-documented in response selection paradigms in general (see e.g., Monsell, 2015, for review) and in language selection in bilinguals in particular (see e.g., Meuter & Allport, 1999). However, what *is* novel in this model is the notion that switch cost does not apply to selecting L1 or L2, but rather switching access to the mental lexicon between comprehension and production processes, regardless of language.

A major simplification in the computational model proposed in Chapter 2 is the absence of contextual facilitation, whether syntactic or semantic. It may seem obvious that in simultaneous interpreting, as in other language tasks, comprehension and production of words that are predictable from context could be facilitated, but since the locus and particularly the magnitude of such facilita-

tory effects cannot be easily derived from the literature, I elected to omit them in the interest of parsimony and see if I could nevertheless obtain a good fit to the observed data. However, while I was able to fit the model's predictions to the observed error rates, this was not direct evidence that my simplifications were warranted.

In Chapter 3 I therefore subjected my simplification assumptions to a stronger test, by removing the narrative context from the stimuli I used in Chapter 2 and presenting participants with randomized word lists to shadow and interpret. If syntactic and semantic context truly had a negligible influence in Chapter 2, the computational model's predictions should have fit behavioral data from word lists just as well as they fit behavioral data from narratives. Instead, I found that the computational model's predictions substantially deviated from the observed error rates unless I made unrealistic assumptions (such as a switch cost in excess of 500 milliseconds for simultaneous interpreting) indicating that the model, as presented in Chapters 2 and 3, does not properly account for facilitation from narrative context. Rather than immediately revisiting the computational model and implementing an ad hoc contextual facilitation "fix", I opted to first examine the loci and magnitude of the unaccounted for contextual facilitation effects. In future work, this information will allow for the making of principled adaptations to the computational model, accounting for contextual facilitation and improving the fit to error rates in the studies reported in this dissertation.

In Chapter 4, examining contextual facilitation, I contrasted semantic, syntactic, and strictly wordform (n-gram based) predictors with the aim of describing at which processing stages contextual facilitation occurs during interpreting and shadowing, and how this affects the difficulty of each task. The resulting linear regression model posed several methodological challenges, both in terms of its technical complexity and in the risk of overfitting caused by the large number of predictors. To mitigate the overfitting problem, I used a relatively novel Bayesian

method where a sparse set of predictors is selected implicitly by setting sparsity-inducing priors on the regression coefficients. The use of this sparse Bayesian regression technique to model data from a psycholinguistic experiment is a novel contribution, but one that has wide-ranging applicability in the field as it strives for increased ecological validity and more naturalistic paradigms, relaxing certain aspects of experimental control and growing the number of potential predictors in each study.

Ultimately, Chapter 4 demonstrated that contextual facilitation does play a clear role in both simultaneous interpreting and in speech shadowing, albeit to a lesser extent in the latter. Fully exploring the consequences of these findings was not possible in the time alotted for this dissertation project, but future iterations of the computational model presented in Chapter 2 should integrate the findings from Chapter 4 in order to more faithfully model the cognitive processes that are essential to interpreting and shadowing, and potentially improve the fit of the model's predictions and the data collected in Chapter 3.

Chapters 5 and 6 did not deal with simultaneous interpreting and speech shadowing directly, instead they describe more general work done in service of the study reported in Chapter 4. Consequently, Chapters 5 and 6 are more methodological in nature and more broadly relevant for psycholinguistics in general. Chapter 5 describes mathematical equivalences between linear combinations of lexical frequency and transitional probability measures, as well as correlations between some of these measures in a large corpus of transcribed speech. I set out to document these issues after finding that they caused multicollinearity in the statistical model reported in Chapter 4, hampering interpretation of the regression coefficients. Any study reporting a linear model with lexical frequency and transitional probability or surprisal measures is susceptible to the problems outlined in Chapter 5, yet psycholinguists publishing these studies generally

make no mention of them when interpreting their results, suggesting they are largely unaware of the extent of the problem.

Finally, in Chapter 6 I present a novel set of word embeddings, a class of distributional semantics models, trained on a corpus of transcribed (pseudoconversational) speech. These models are commonly trained on written text from a particular register (e.g., a newspaper corpus or the entirety of Wikipedia, as in Grave et al., 2018) which makes their validity for modeling speech data questionable. I systematically evaluated the difference between word embeddings trained on Wikipedia and transcribed speech, ultimately finding smaller differences than expected, but highlighting the benefits of using diverse training corpora when training distributional semantics models. Chapter 6 furthermore introduces novel evaluation metrics for word embeddings, facilitating systematic benchmarking of future efforts in distributional semantics for psycholinguistics. All code and data can be downloaded and reproduced conveniently using a publicly available Python software package, making this a significant novel resource for psycholinguists interested in quantifying semantics, especially in speech (e.g., when statistically modeling semantic priming data). The datasets compiled in Chapter 6 were essential to quantifying the degree of semantic facilitation in the model reported in Chapter 4.

7.2 Speaking and listening in simultaneous interpreting and speech shadowing

The first half of this dissertation represents an attempt to explain the aspect of simultaneous interpreting that I find most striking: It seems like it should not be possible to speak and listen at the same time, and yet professional interpreters seem to do it quite fluently. Formulating and producing speech requires the use of lemma and conceptual-semantic networks, and yet at the time that these

networks should be activated in simultaneous interpreting, more speech stimuli need to be processed on the comprehension side, which presumably also requires the use of lemma and conceptual-semantic networks. In this dissertation I have shown that error rates from a simultaneous interpreting and speech shadowing experiment are broadly consistent with a model where access to lemma selection and retrieval is switched between comprehension and production processes, incurring a small switch cost on every switch. This gives the appearance of concurrent speaking and listening at slow speech rates, since there is some buffering on the sensory and articulation ends, and the natural pauses in speech can absorb some of the switch costs. At high speech rates, however, the switch costs can no longer be absorbed and lapses in comprehension and production become apparent, giving the impression of alternating speaking and listening.

In reality, speaking and listening are neither fully concurrent nor fully alternating: low-level processes that do not draw on networks shared between speaking and listening can occur concurrently, while higher-level processes that draw on networks shared between speaking and listening must alternate. This model obviates the need for separate input and output lexicons, as posited in Christoffels and De Groot (2004), which as discussed in Chapter 1 is not very plausible, since it is unclear how novel words would enter into the production lexicon from the comprehension lexicon.

The model also has implications for speaking and listening in conversation: If while listening, we are unable to fully plan and formulate what we are going to say when it is our turn to speak, this suggests that rapid turn-taking in conversation entails less than optimal listening (as well as less than optimal speech planning, potentially) and perhaps no real, attentive listening at all. If we are to respond quickly to an utterance once our interlocutor stops speaking, one strategy is to stop attending to their speech once we have understood their message, yet well before they stop speaking, so we can plan our utterance while our in-

terlocutor's speech goes unattended. There is some experimental evidence that people indeed plan their speech in this manner, while relying on other cues to determine when their interlocutor is done speaking and they can initiate their own turn (Barthel et al., 2017).

Concurrency and switch cost aside, speech comprehension and production in simultaneous interpreting are not so different from normal speaking and listening, as evidenced by my findings in Chapter 4. I find semantic priming, cognate priming, word length and frequency effects, and effects from syntactic processing load, all broadly consistent with effects observed in normal speaking and listening. Speech shadowing does not exhibit the semantic priming and syntactic processing effects, only the word-level facilitation effects, consistent with the idea that shadowing is conducted primarily along a lower-level route than simultaneous interpreting and is in that sense much less like normal speaking or listening. Marslen-Wilson (1985) reported that shadowers with a lower shadowing latency were less able to detect semantic mismatches in attended speech, convergent evidence that close shadowing largely bypasses conceptual-semantic understanding in favor of a lower-level route from comprehension to production.

It should be noted however that we cannot easily compare the effect sizes obtained in Chapter 4 to those observed in the psycholinguistic literature, because most studies examining these effects have been more neatly controlled experiments using either single-word production paradigms or various comprehension paradigms that do not include a production component. However, that I was able to observe these effects in continuous speech (and under the high cognitive load associated with shadowing and interpreting) suggests that they are fairly robust to variations in experimental paradigms and task demands.

7.3 Studying speech comprehension and speech production concurrently

In general, in the behavioral domain speech comprehension and speech production have traditionally been studied separately, and in the case of speech production, predominantly at the level of single words. Both in terms of experimental design and statistical methods it is easier to conduct experiments at the single word level, and because there are only lexical factors to consider, it is often conceptually clearer. This approach has yielded a good picture of the time course of both single word comprehension and production (see e.g., Indefrey & Levelt, 2004; Indefrey, 2011). However, the time course of speech comprehension and production processes for larger linguistic structures is less settled.

While there are theories of speech production that make predictions at a larger scope than the lexical (or noun-phrase) scope, the difficulty of reliably eliciting planning and production of sentences or narratives (as opposed to simple sentence reading, for instance) makes it difficult to test these theories. The twin tasks of speech shadowing and simultaneous interpreting present us with a paradigm to elicit concurrent speech comprehension and production processes. We can derive an online measure of difficulty of the task in the form of speech production error rates (requiring translation, in the case of interpreting) and input-output speech latencies. Online measures of speech comprehension and production difficulty are generally based on psychophysiological or neural correlates (eye movements, EEG, MEG, and sometimes pupillometry) so having a direct behavioral measure presents an unusual opportunity for investigating concurrence in speech comprehension and production.

7.4 Comparing the model to its predecessors

In terms of complexity, the model proposed in Chapter 2 falls somewhere between the process models from the 1970s (e.g., Gerver, 1975; Moser, 1978), which were simply too richly detailed to ever implement computationally or to otherwise glean quantitative predictions from, and a few more recent process models which aim to explain aspects of interpreting such as attentional control (Dong & Li, 2019) and prediction (Amos & Pickering, 2020) and which are less overdetailed, but so general in their claims as to be virtually unfalsifiable.

In spirit, the model proposed in this dissertation most closely resembles a process model as presented by Christoffels and De Groot (2004), except implementing it computationally allows for testing agains behavioral data. More specifically, in this dissertation the implemented process model was tested against interpreting and shadowing error rates, but naturally the speech latencies that were (quite laboriously) annotated later on the same data for the purpose of conducting the study reported Chapter 4 present an additional opportunity to test the model. I have yet to adapt the computational implementation of the model to test its predictions against these speech latencies, but this represents an interesting avenue for future exploration. A reimplementation of the process model could also be designed to account for the contextual facilitation effects identified in Chapter 4, although how to model contextual facilitation in a manner both parsimonious and psycholinguistically plausible is not immediately obvious (i.e., as discussed previously, simply adding a “contextual facilitation” module would not be satisfactory).

7.5 Wrongness and abstraction in computational and descriptive models

George Box's aphorism that "all models are wrong, but some are useful" often comes up in discussions about computational modeling, especially when a model fails to capture an edge case, or modeling results are found not to generalize as one had hoped (as in Chapter 3 of this dissertation, for instance). While a model being potentially useful even when demonstrably wrong to some extent is a comforting thought, it is worth noting that Box generally expressed more open-ended versions of this sentiment in his writings, e.g., "Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful." (Box & Draper, 1987, p. 74). This version makes it clearer that utility in the face of wrongness is not a given and "wrong" here does not refer to deliberate wrongheadedness, but to the necessity for models to contain abstractions, and as a result not to capture every aspect of the phenomenon they are meant to represent; "A map is not the territory it represents, but, if correct, it has a similar structure to the territory, which accounts for its usefulness." (Korzybski, 1933, p. 58).

There is a tendency in cognitive modeling attempt to increase a model's explanatory power by making it more comprehensive, capturing more edge cases, more nuances, more behaviors. This approach is seductive, but it seldomly works well, and ultimately models where the "map has become the territory" are invariably too complex to actually explain the phenomena they model in a manner that improves understanding (see e.g., recent advances in language modeling with GPT-3, a model that exhibits astounding performance on various language tasks, but contains 175 billion parameters).

Finally, Box's aphorism is recited almost as a mantra by some computational modelers, but is often forgotten in the context of descriptive models, for which

the implicit assumption is often that whatever is not explicitly modeled does not materially affect the phenomenon that is being modeled. When modeling a behavior as complex as language, descriptive models tend to oversimplify where computational models might overcomplicate. In Chapter 4 of this dissertation I hope to have gotten closer to the territory than descriptive models of language comprehension and production generally do, by modeling a wide array of predictors and their effects on interpreting and shadowing latencies.

7.6 Conclusion

Ultimately, while the model proposed in Chapter 2 does not generalize well to interpreting in the absence of narrative context, the key assumption introduced in the model—that speech comprehension and production processes can be largely independent but at a minimum need to alternate access to selection processes in the mental lexicon so as not to catastrophically interfere—is consistent with both patterns of behavior and error in interpreting narratives and the broader psycholinguistic literature. It is clear from Chapter 3 that accounting for facilitation from narrative context is crucial in order to achieve further progress in the computational modeling of complex language tasks such as simultaneous interpreting and speech shadowing; Chapter 4 demonstrates several types of contextual facilitation as they occur in interpreting and shadowing and their relative magnitude, potential starting values for amending the computational model introduced in Chapter 2.

A more general contribution of this dissertation (and perhaps more widely applicable) is the first demonstration, in Chapter 4, of how to apply Bayesian sparse regression techniques to a complex, naturalistic language task that poses a problem for conventional linear regression techniques. While a number of technical and computational problems will require further work, the potential of this tech-

nique to address foundational methodological issues in psycholinguistics is clear. Combined with the semantic resources introduced in Chapter 6 and the mathematical considerations presented in Chapter 5, the latter half of this dissertation offers a roadmap for improved descriptive modeling of complex language tasks.

References

Abella, R. A. S. M., & González-Nosti, M. (2019). Motor content norms for 4,565 verbs in Spanish. *Behavior Research Methods*, 1–8. https://doi.org/10.3758/s13428-019-01241-1

Alday, P. M. (2019). M/EEG analysis of naturalistic stories: A review from speech to language processing. *Language, Cognition and Neuroscience*, *34*(4), 457–473. https://doi.org/10.1080/23273798.2018.1546882

Alday, P. M., Schlesewsky, M., & Bornkessel-Schlesewsky, I. (2017). Electrophysiology reveals the neural dynamics of naturalistic auditory language processing: Event-related potentials reflect continuous model updates. *eNeuro*, *4*(6). https://doi.org/10.1523/ENEURO.0311-16.2017

Al-Rfou, R., Perozzi, B., & Skiena, S. (2013). Polyglot: Distributed word representations for multilingual NLP. http://arxiv.org/abs/1307.1662

Altmann, G. T., & Mirkovi, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*, *33*(4), 583–609.

Amos, R. M., & Pickering, M. J. (2020). A theory of prediction in simultaneous interpreting. *Bilingualism: Language and Cognition*, 1–10.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*(1), 67–82. https://doi.org/10.1016/j.jml.2009.09.005

Audacity Team. (2017). Audacity: Free audio editor and recorder (version 2.0.6)[computer software].

Baayen, R. H., Hendrix, P., & Ramscar, M. (2013). Sidestepping the combinatorial explosion: An explanation of n-gram frequency effects based on naive discriminative learning. *Language and Speech*. https://doi.org/10.1177/0023830913484896

Baker, S., Reichart, R., & Korhonen, A. (2014). An unsupervised model for instance level subcategorization acquisition. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 278–289.

Bakhtiar, M., & Weekes, B. (2015). Lexico-semantic effects on word naming in Persian: does age of acquisition have an effect? *Memory and Cognition*, *43*, 298–313. https://doi.org/10.3758/s13421-014-0472-4

Balota, D. A., Pollatsek, A., & Rayner, K. (1985). The interaction of contextual constraints and parafoveal visual information in reading. *Cognitive Psychology*, *17*(3), 364–390. https://doi.org/10.1016/0010-0285(85)90013-1

Barik, H. C. (1973). Simultaneous interpretation: Temporal and quantitative data. *Language and speech*, *16*(3), 237–270.

Barthel, M., Meyer, A. S., & Levinson, S. C. (2017). Next speakers plan their turn early and speak after turn-final go-signals. *Frontiers in Psychology*, *8*, 393. https://doi.org/10.3389/fpsyg.2017.00393

Bates, D., Maechler, M., Bolker, B., & Walker., S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Berardi, G., Esuli, A., & Marcheggiani, D. (2015). Word embeddings go to Italy: A comparison of models and training datasets. *Proceedings of the Italian Information Retrieval Workshop*.

Bestgen, Y. (2008). Building affective lexicons from specific corpora for automatic sentiment analysis. In N. Calzolari, K. Choukri, B.Maegaard, J. Mariani, J. Odjik, S. Piperidis, & D. Tapias (Eds.), *Proceedings of lrec '08, 6th language resources and evaluation conference* (pp. 496–500). ELRA.

Bestgen, Y., & Vincze, N. (2012). Checking and bootstrapping lexical norms by means of word similarity indexes. *Behavior Research Methods*, *44*(4), 998–1006. https://doi.org/10.3758/s13428-012-0195-z

Betancourt, M. (2018). Bayes Sparse Regression.

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. (2016). Toward a brain-based componential semantic representation. *Cognitive Neuropsychology*, *33*(3-4), 130–174. https://doi.org/10.1080/02643294.2016.1147426

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *arXiv:1607.04606 [cs]*.

Bonin, P., Méot, A., & Bugaiska, A. (2018). Concreteness norms for 1,659 French words: Relationships with other psycholinguistic variables and word recognition times. *Behavior Research Methods*, *50*(6), 2366–2387. https://doi.org/10.3758/s13428-018-1014-y

Box, G. E., & Draper, N. R. (1987). *Empirical model-building and response surfaces*. John Wiley & Sons.

Broadbent, D. E. (1967). Word-frequency effect and response bias. *Psychological Review*, *74*(1), 1. https://doi.org/10.1037/h0024206

Broadbent, D. E. (1952). Speaking and listening simultaneously. *Journal of experimental psychology*, *43*(4), 267.

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, *28*(5), 803–809.e3. https://doi.org/10.1016/j.cub.2018.01.080

Broderick, M. P., Anderson, A. J., & Lalor, E. C. (2019). Semantic context enhances the early auditory encoding of natural speech. *Journal of Neuroscience*, *39*(38), 7564–7575. https://doi.org/10.1523/JNEUROSCI.0584-19.2019

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178.

Bruni, E., Boleda, G., Baroni, M., & Tran, N.-K. (2012). Distributional semantics in technicolor. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, 136–145.

Brysbaert, M., Keuleers, E., & New, B. (2011). Assessing the usefulness of Google Books word frequencies for psycholinguistic research on word processing. *Frontiers in Psychology*, *2*, 27. https://doi.org/10.3389/fpsyg.2011.00027

Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2019). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*, *51*(2), 467–479. https://doi.org/10.3758/s13428-018-1077-9

Brysbaert, M., & New, B. (2009). Moving beyond Kuera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., Stevens, M., De Deyne, S., Voorspoels, W., & Storms, G. (2014). Norms of age of acquisition and concreteness for 30,000 Dutch words. *Acta Psychologica*, *150*, 80–84. https://doi.org/10.1016/j.actpsy.2014.04.010

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Cameirão, M. L., & Vicente, S. G. (2010). Age-of-acquisition norms for a set of 1,749 Portuguese words. *Behavior Research Methods*, *42*(2), 474–480. https://doi.org/10.3758/BRM.42.2.474

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, *97*(2), 465–480. https://doi.org/10.1093/biomet/asq017

Chedid, G., Brambati, S. M., Bedetti, C., Rey, A. E., Wilson, M. A., & Vallet, G. T. (2019). Visual and auditory perceptual strength norms for 3,596 French nouns and their relationship with other psycholinguistic variables. *Behavior Research Methods*, *51*(5), 2094–2105. https://doi.org/10.3758/s13428-019-01254-w

Chedid, G., Wilson, M. A., Bedetti, C., Rey, A. E., Vallet, G. T., & Brambati, S. M. (2019). Norms of conceptual familiarity for 3,596 French nouns and their contribution in lexical decision. *Behavior Research Methods*, *51*(5), 2238–2247. https://doi.org/10.3758/s13428-018-1106-8

Chen, D., Peterson, J. C., & Griffiths, T. L. (2017). Evaluating vector-space models of analogy. *arXiv preprint arXiv:1705.04416*.

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*.

Christoffels, I. K., & De Groot, A. M. B. (2004). Components of simultaneous interpreting: Comparing interpreting with shadowing and paraphrasing. *Bilingualism: Language and Cognition*, *7*(3), 227–240. https://doi.org/10.1017/S1366728904001609

Christoffels, I. K., De Groot, A. M. B., & Waldorp, L. J. (2003). Basic skills in a complex task: A graphical model relating memory and lexical retrieval to simultaneous interpreting. *Bilingualism: Language and Cognition*, *6*(3), 201–211. https://doi.org/10.1017/S1366728903001135

Christoffels, I. K., De Groot, A. M., & Kroll, J. F. (2006). Memory and language skills in simultaneous interpreters: The role of expertise and language proficiency. *Journal of Memory and Language*, *54*(3), 324–345.

Christoffels, I. K., Firk, C., & Schiller, N. O. (2007). Bilingual language control: An event-related brain potential study. *Brain Research*, *1147*, 192–208. https://doi.org/10.1016/j.brainres.2007.01.137

Claesen, M., Simm, J., Popovic, D., Moreau, Y., & De Moor, B. (2014). Easy hyperparameter search using optunity. *arXiv preprint arXiv:1412.1114*.

Cleland, A. A., Gaskell, M. G., Quinlan, P. T., & Tamminen, J. (2006). Frequency effects in spoken and visual word recognition: Evidence from dual-task methodologies. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(1), 104. https://doi.org/10.1037/0096-1523.32.1.104

Cook, A. E., & Meyer, A. S. (2008). Capacity demands of phoneme selection in word production: New evidence from dual-task experiments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(4), 886.

Costa, A., Caramazza, A., & Sebastian-Galles, N. (2000). The cognate facilitation effect: Implications for models of lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1283–1296. https://doi.org/10.1037/0278-7393.26.5.1283

Costa, A., Santesteban, M., & Caño, A. (2005). On the facilitatory effects of cognate words in bilingual speech production. *Brain and Language*, *94*(1), 94–103. https://doi.org/10.1016/j.bandl.2004.12.002

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time Course of Frequency Effects in Spoken-Word Recognition: Evidence from Eye Movements. *Cognitive Psychology*, *42*(4), 317–367. https://doi.org/10.1006/cogp.2001.0750

De Groot, A. M. (1992). Determinants of word translation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(5), 1001.

Desrochers, A., & Thompson, G. L. (2009). Subjective frequency and imageability ratings for 3,600 french nouns. *Behavior Research Methods*, *41*(2), 546–557. https://doi.org/10.3758/BRM.41.2.546

Díez-Álamo, A. M., Díez, E., Alonso, M. Á., Vargas, C. A., & Fernandez, A. (2018). Normative ratings for perceptual and motor attributes of 750 object concepts in Spanish. *Behavior Research Methods*, *50*(4), 1632–1644. https://doi.org/10.3758/s13428-017-0970-y

Díez-Álamo, A. M., Díez, E., Wojcik, D. Z., Alonso, M. A., & Fernandez, A. (2019). Sensory experience ratings for 5,500 Spanish words. *Behavior Research Methods*, *51*(3), 1205–1215. https://doi.org/10.3758/s13428-018-1057-0

Dijkstra, T., Wahl, A., Buytenhuijs, F., Van Halem, N., Al-Jibouri, Z., De Korte, M., & Rekké, S. (2019). Multilink: A computational model for bilingual word recognition and word translation. *Bilingualism: Language and Cognition*, *22*(4), 657–679.

Dong, Y., & Li, P. (2019). Attentional control in interpreting: A model of language control and processing control. *Bilingualism: Language and Cognition*, 1–13.

Dos Santos, L. B., Duran, M. S., Hartmann, N. S., Candido, A., Paetzold, G. H., & Aluisio, S. M. (2017). A lightweight regression method to infer psycholinguistic properties for Brazilian Portuguese. *International Conference on Text, Speech, and Dialogue*, 281–289.

Eilola, T. M., & Havelka, J. (2010). Affective norms for 210 British English and Finnish nouns. *Behavior Research Methods*, *42*(1), 134–140. https://doi.org/10.3758/BRM.42.1.134

Engelthaler, T., & Hills, T. T. (2018). Humor norms for 4,997 English words. *Behavior Research Methods*, *50*(3), 1116–1124. https://doi.org/10.3758/s13428-017-0930-6

Faruqui, M., Tsvetkov, Y., Rastogi, P., & Dyer, C. (2016). *Problems with evaluation of word embeddings using word similarity tasks*. arXiv: 1605.02276.

Feng, S., Cai, Z., Crossley, S. A., & McNamara, D. S. (2011). Simulating human ratings on word concreteness. *FLAIRS Conference*.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., Augustinova, M., & Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 french words and 38,840 pseudowords. *Behavior Research Methods*, *42*(2), 488–496. https://doi.org/10.3758/BRM.42.2.488

Ferreira, V. S., & Pashler, H. (2002). Central bottleneck influences on the processing stages of word production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(6), 1187.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin., E. (2001). Placing search in context: The concept revisited. *Proceedings of the 10th International Conference on World Wide Web*. https://doi.org/10.1145/503104.503110

Fowler, C. A., Brown, J. M., Sabadini, L., & Weihing, J. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory and Language*, *49*(3), 396–413.

Frank, S. L. (2013). Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in Cognitive Science*, *5*(3), 475–494. https://doi.org/10.1111/tops.12025

Frank, S. L., & Willems, R. M. (2017). Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience*, *32*(9), 1192–1203. https://doi.org/10.1080/23273798.2017.1323109

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

Friston, K., Thornton, C., & Clark, A. (2012). Free-Energy Minimization and the Dark-Room Problem. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00130

Gaiba, F. (1998). *The origins of simultaneous interpretation: The nuremberg trial*. University of Ottawa Press.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644.

Gerver, D. (1969). The effects of source language presentation rate on the performance of simultaneous conference interpreters. *Proceedings of the 2nd Louisville Conference on Rate and/or Frequency Controlled Speech*, 162–184.

Gerver, D. (1974a). The effects of noise on the performance of simultaneous interpreters: Accuracy of performance. *Acta Psychologica*, *38*(3), 159–167.

Gerver, D. (1974b). Simultaneous listening and speaking and retention of prose. *The Quarterly journal of experimental psychology*, *26*(3), 337–341.

Gerver, D. (1975). A psychological approach to simultaneous interpretation. *Meta: Journal des traducteurs/Meta: Translators' Journal*, *20*(2), 119–128.

Gerz, D., Vulic, I., Hill, F., Reichart, R., & Korhonen, A. (2016). SimVerb-3500: a large-scale evaluation set of verb similarity.

Gile, D., Danks, H., Shreve, G., Fountain, S., & McBeath, M. (1997). Cognitive processes in translation and interpreting. *Conference Interpreting as a Cognitive Management Problem, edited by Joseph E. Danks, Gregory M. Shreve, Stephen B. Fountain, and Michael K. McBeath*, 196–214.

Goldman-Eisler, F. (1972). Segmentation of input in simultaneous translation. *Journal of psycholinguistic Research*, *1*(2), 127–140.

Göz,., Tekcan, A.., & Erciyes, A. A. (2017). Subjective age-of-acquisition norms for 600 Turkish words from four age groups. *Behavior Research Methods*, *49*(5), 1736–1746. https://doi.org/10.3758/s13428-016-0817-y

Grandy, T. H., Lindenberger, U., & Schmiedek, F. (2020). Vampires and nurses are rated differently by younger and older adults–Age-comparative norms of imageability and emotionality for about 2500 German nouns. *Behavior Research Methods*, 1–10. https://doi.org/10.3758/s13428-019-01294-2

Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, *37*(3), 424–438. https://doi.org/10.2307/1912791

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). *Learning word vectors for 157 languages*. arXiv: 1802.06893.

Guasch, M., Ferré, P., & Fraga, I. (2016). Spanish norms for affective and lexico-semantic variables for 1,400 words. *Behavior Research Methods*, *48*(4), 1358–1369. https://doi.org/10.3758/s13428-015-0684-y

Gurevych, I. (2005). Using the structure of a conceptual network in computing semantic relatedness. *In Proceedings of the International Joint Conference on Natural Language Processing*. https://doi.org/10.1007/11562214_67

Halawi, G., Dror, G., Gabrilovich, E., & Koren, Y. (2012). Large-scale learning of word relatedness with constraints. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1406–1414. https://doi.org/10.1145/2339530.2339751

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

Hassan, S., & Mihalcea, R. (2009). Cross-lingual semantic relatedness using encyclopedic knowledge. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Hervais-Adelman, A., Moser-Mercer, B., & Golestani, N. (2015). Brain functional plasticity associated with the emergence of expertise in extreme language control. *NeuroImage*, *114*, 264–274.

Hervais-Adelman, A., Moser-Mercer, B., Michel, C. M., & Golestani, N. (2015). Fmri of simultaneous interpretation reveals the neural basis of extreme language control. *Cerebral Cortex*, *25*(12), 4727–4739. https://doi.org/10.1093/cercor/bhu158

Hill, F., Reichart, R., & Korhonen, A. (2014). SimLex-999: evaluating semantic models with (genuine) similarity estimation. *Computing Research Repository*.

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin & Review*, *23*(6), 1744–1756. https://doi.org/10.3758/s13423-016-1053-2

Hollis, G., Westbury, C., & Lefsrud, L. (2017). Extrapolating human judgments from skip-gram vector representations of word meaning. *The Quarterly Journal of Experimental Psychology*, *70*(8), 1603–1619. https://doi.org/10.1080/17470218.2016.1195417

Honnibal, M., & Johnson, M. (2015). An Improved Non-monotonic Transition System for Dependency Parsing. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1373–1378. https://doi.org/10.18653/v1/D15-1162

Imbir, K. K. (2015). Affective norms for 1,586 Polish words (ANPW): Duality-of-mind approach. *Behavior Research Methods*, *47*(3), 860–870. https://doi.org/10.3758/s13428-014-0509-4

Indefrey, P., & Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. *Cognition*, *92*(1), 101–144. https://doi.org/10.1016/j.cognition.2002.06.001

Indefrey, P. (2011). The spatial and temporal signatures of word production components: A critical update. *Frontiers in Psychology*, *2*, 255. https://doi.org/10.3389/fpsyg.2011.00255

Jacobs, C. L., & Dell, G. S. (2014). 'hotdog', not 'hot' 'dog': The phonological planning of compound words. *Language, Cognition and Neuroscience*, *29*(4), 512–523. https://doi.org/10.1080/23273798.2014.892144

Jacobs, C. L., Dell, G. S., Benjamin, A. S., & Bannard, C. (2016). Part and whole linguistic experience affect recognition memory for multiword sequences. *Journal of Memory and Language*, *87*, 38–58. https://doi.org/10.1016/j.jml.2015.11.001

Janschewitz, K. (2008). Taboo, emotionally valenced, and emotionally neutral word norms. *Behavior Research Methods*, *40*(4), 1065–1074. https://doi.org/10.3758/BRM.40.4.1065

Janssen, N., & Barber, H. A. (2012). Phrase Frequency Effects in Language Production. *PLOS ONE*, *7*(3), e33202. https://doi.org/10.1371/journal.pone.0033202

Jescheniak, J. D., & Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 824–843. https://doi.org/10.1037/0278-7393.20.4.824

Jescheniak, J. D., Meyer, A. S., & Levelt, W. J. M. (2003). Specific-word frequency is not all that counts in speech production: Comments on Caramazza, Costa, et al. (2001) and new experimental data. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(3), 432–438. https://doi.org/10.1037/0278-7393.29.3.432

Joubarne, C., & Inkpen, D. (2011). Comparison of semantic similarity for different languages using the Google n-gram corpus and second-order

co-occurrence measures. *Proceedings of the Canadian Conference on Artificial Intelligence*. https://doi.org/10.1007/978-3-642-21043-3_26

Kanske, P., & Kotz, S. A. (2010). Leipzig Affective Norms for German: A reliability study. *Behavior Research Methods*, *42*(4), 987–991. https://doi.org/10.3758/BRM.42.4.987

Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. *Proceedings of ICNN'95-International Conference on Neural Networks*, *4*, 1942–1948. https://doi.org/10.1109/ICNN.1995.488968

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: a new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, *42*(3), 643–650. https://doi.org/10.3758/BRM.42.3.643

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The british lexicon project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*(1), 287–304. https://doi.org/10.3758/s13428-011-0118-4

Keuleers, E., Stevens, M., Mandera, P., & Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *The Quarterly Journal of Experimental Psychology*, *68*(8), 1665–1692. https://doi.org/10.1080/17470218.2015.1022560

Kisler, T., Reichel, U., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, *45*, 326–347. https://doi.org/10.1016/j.csl.2017.01.005

Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, *16*(1-2), 262–284. https://doi.org/10.1080/09541440340000213

Konopka, A. E., & Meyer, A. S. (2014). Priming sentence planning. *Cognitive Psychology*, *73*, 1–40. https://doi.org/10.1016/j.cogpsych.2014.04.001

Köper, M., Scheible, C., & im Walde, S. S. (2015). Multilingual reliability and semantic structure of continuous word spaces. *Proceedings of the International Conference on Computational Semantics*.

Korzybski, A. (1933). Science and sanity. an introduction to non-aristotelian systems and general semantics.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., & Blei, D. M. (2017). Automatic differentiation variational inference. *The Journal of Machine Learning Research*, *18*(1), 430–474.

Kuperman, V. (2013). Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers. https://doi.org/10.1037/a0030859

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

Kuznetsova, A., Brockhoff, P. B., Christensen, R. H., et al. (2017). Lmertest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

La Heij, W., Hooglander, A., Kerling, R., & Van Der Velden, E. (1996). Nonverbal context effects in forward and backward word translation: Evidence for concept mediation. *Journal of Memory and Language*, *35*(5), 648–665. https://doi.org/10.1006/jmla.1996.0034

Lehtonen, M., Soveri, A., Laine, A., Järvenpää, J., De Bruin, A., & Antfolk, J. (2018). Is bilingualism associated with enhanced executive functioning in adults? a meta-analytic review. *Psychological bulletin*, *144*(4), 394.

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, *44*(2), 325–343. https://doi.org/10.3758/s13428-011-0146-0

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. The MIT Press.

Levelt, W. J. M., & Wheeldon, L. (1994). Do speakers have access to a mental syllabary? *Cognition*, *50*(1), 239–269. https://doi.org/10.1016/0010-0277(94)90030-2

Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1–38. https://doi.org/10.1017/S0140525X99001776

Levinson, S. C. (2016). Turn-taking in human communication–origins and implications for language processing. *Trends in cognitive sciences*, *20*(1), 6–14.

Levy, O., & Goldberg, Y. (2014). Linguistic regularities in sparse and explicit word representations. *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, 171–180. https://doi.org/10.3115/v1/W14-1618

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Lewis, M., Zettersten, M., & Lupyan, G. (2019). Distributional semantics as a source of visual knowledge. *Proceedings of the National Academy of Sciences*, *116*(39), 19237–19238. https://doi.org/10.1073/pnas.1910148116

Liang, J., Fang, Y., Lv, Q., & Liu, H. (2017). Dependency Distance Differences across Interpreting Types: Implications for Cognitive Demand. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.02132

Luong, T., Socher, R., & Manning, C. (2013). Better word representations with recursive neural networks for morphology. *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, 104–113.

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2019). The Lancaster Sensorimotor Norms: multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, 1–21. https://doi.org/10.3758/s13428-019-01316-z

Mandera, P., Keuleers, E., & Brysbaert, M. (2015). How useful are corpus-based methods for extrapolating psycholinguistic variables? *The Quarterly Journal of Experimental Psychology*, *68*(8), 1623–1642. https://doi.org/10.1080/17470218.2014.988735

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78. https://doi.org/10.1016/j.jml.2016.04.001

Marslen-Wilson, W. D. (1973). Linguistic structure and speech shadowing at very short latencies. *Nature*, *244*(5417), 522–523.

Marslen-Wilson, W. D. (1975). Sentence perception as an interactive parallel process. *Science*, *189*(4198), 226–228.

Marslen-Wilson, W. D. (1985). Speech shadowing and speech comprehension. *Speech Communication*, *4*(1-3), 55–73.

McDonald, S. A., & Shillcock, R. C. (2003). Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science*, *14*(6), 648–652. https://doi.org/10.1046/j.0956-7976.2003.psci_1480.x

Meuter, R. F., & Allport, A. (1999). Bilingual language switching in naming: Asymmetrical costs of language selection. *Journal of Memory and Language*, *40*(1), 25–40. https://doi.org/10.1006/jmla.1998.2602

Meyer, A. S. (1990). The time course of phonological encoding in language production: The encoding of successive syllables of a word. *Journal of Memory and Language*, *29*(5), 524–545. https://doi.org/10.1016/0749-596X(90)90050-A

Meyer, A. S., Roelofs, A., & Levelt, W. J. M. (2003). Word length effects in object naming: The role of a response criterion. *Journal of Memory and Lan-*

guage, *48*(1), 131–147. https://doi.org/10.1016/S0749-596X(02)00509-0

Meyer, A. S., Sleiderink, A. M., & Levelt, W. J. M. (1998). Viewing and naming objects: Eye movements during noun phrase production. *Cognition*, *66*(2), B25–B33. https://doi.org/10.1016/S0010-0277(98)00009-2

Meyer, C. M., & Gurevych, I. (2012). To exhibit is not to loiter: A multilingual, sense-disambiguated wiktionary for measuring verb similarity. *Proceedings of COLING 2012*, 1763–1780.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv: 1301.3781.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2017). *Advances in pre-training distributed word representations*. arXiv: 1712.09405.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality*. arXiv: 1310.4546.

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, *4*(1), 1–28. https://doi.org/10.1080/01690969108406936

Mizuno, A. (2005). Process model for simultaneous interpreting and working memory. *Meta: Journal des Traducteurs/Meta: Translators' Journal*, *50*(2), 739–752.

Monnier, C., & Syssau, A. (2014). Affective norms for French words (FAN). *Behavior Research Methods*, *46*(4), 1128–1137. https://doi.org/10.3758/s13428-013-0431-1

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134–140.

Monsell, S. (2015). Task-set control and task switching. In J. M. Fawcett, E. F. Risko, & A. Kingstone (Eds.), *The handbook of attention* (pp. 139–172). MIT Press Cambridge, MA.

Morales, J., Padilla, F., Gómez-Ariza, C. J., & Bajo, M. T. (2015). Simultaneous interpretation selectively influences working memory and attentional networks. *Acta Psychologica*, *155*, 82–91.

Moser, B. (1978). Simultaneous Interpretation: A Hypothetical Model and its Practical Application. In D. Gerver & H. W. Sinaiko (Eds.), *Language Interpretation and Communication* (pp. 353–368). Springer US. https://doi.org/10.1007/978-1-4615-9077-4_31

Nathanson, A. I., Aladé, F., Sharp, M. L., Rasmussen, E. E., & Christy, K. (2014). The relation between television exposure and executive function among

preschoolers. *Developmental Psychology*, *50*(5), 1497. https://doi.org/10.1037/a0035714

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied psycholinguistics*, *28*(4), 661–677. https://doi.org/10.1017/S014271640707035X

Oldfield, R. C., & Wingfield, A. (1965). Response Latencies in Naming Objects: *Quarterly Journal of Experimental Psychology*. https://doi.org/10.1080/17470216508416445

Ostarek, M., Van Paridon, J., & Montero-Melis, G. (2019). Sighted peoples language is not helpful for blind individuals' acquisition of typical animal colors. *Proceedings of the National Academy of Sciences*, *116*(44), 21972–21973. https://doi.org/10.1073/pnas.1912302116

Paap, K., Schwieter, J., & Paradis, M. (2019). The bilingual advantage debate. *The handbook of the neuroscience of multilingualism* (pp. 701–735). Wiley Online Library.

Panchenko, A., Ustalov, D., Arefyev, N., Paperno, D., Konstantinova, N., Loukachevitch, N., & Biemann, C. (2016). Human and machine judgements for Russian semantic relatedness. *Proceedings of the International Conference, Analysis of Images, Social networks and Texts*. https://doi.org/10.1007/978-3-319-52920-2_21

Pereira, F., Gershman, S., Ritter, S., & Botvinick, M. (2016). A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive Neuropsychology*, *33*(3), 175–190. https://doi.org/10.1080/02643294.2016.1176907

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., Botvinick, M., & Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, *9*(963). https://doi.org/10.1038/s41467-018-03068-4

Pexman, P. M., Muraki, E., Sidhu, D. M., Siakaluk, P. D., & Yap, M. J. (2019). Quantifying sensorimotor experience: Body–object interaction ratings for more than 9,000 English words. *Behavior Research Methods*, *51*(2), 453–466. https://doi.org/10.3758/s13428-018-1171-z

Piai, V., Roelofs, A., & Schriefers, H. (2014). Locus of semantic interference in picture naming: Evidence from dual-task performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(1), 147.

Piironen, J., & Vehtari, A. (2017a). Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, *27*(3), 711–735. https://doi.org/10.1007/s11222-016-9649-y

Piironen, J., & Vehtari, A. (2017b). On the Hyperprior Choice for the Global Shrinkage Parameter in the Horseshoe Prior. *arXiv:1610.05559 [stat]*.

Piironen, J., & Vehtari, A. (2017c). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, *11*(2), 5018–5051. https://doi.org/10.1214/17-EJS1337SI

Postma, M., & Vossen, P. (2014). What implementation and translation teach us: The case of semantic similarity measures in wordnets. *Proceedings of the Seventh Global Wordnet Conference*, 133–141.

Querido, A., de Carvalho, R., Garcia, M., Correia, C., Rendeiro, N., Pereira, R., Campos, M., Branco, A., et al. (2017). LX-LR4DistSemEval: a collection of language resources for the evaluation of distributional semantic models of Portuguese. *Revista da Associação Portuguesa de Linguística*, (3), 265–283.

R Core Team. (2013). R: A language and environment for statistical computing.

Radeau, M., & Morais, J. (1990). The uniqueness point effect in the shadowing of spoken words. *Speech Communication*, *9*(2), 155–164. https://doi.org/10.1016/0167-6393(90)90068-K

Radeau, M., Morais, J., Mousty, P., & Bertelson, P. (2000). The Effect of Speaking Rate on the Role of the Uniqueness Point in Spoken Word Recognition. *Journal of Memory and Language*, *42*(3), 406–422. https://doi.org/10.1006/jmla.1999.2682

Radinsky, K., Agichtein, E., Gabrilovich, E., & Markovitch, S. (2011). A word at a time: Computing word relatedness using temporal semantic analysis. *Proceedings of the 20th international conference on World wide web*, 337–346. https://doi.org/10.1145/1963405.1963455

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372. https://doi.org/10.1037/0033-2909.124.3.372

Recchia, G., & Louwerse, M. M. (2015a). Reproducing affective norms with lexical co-occurrence statistics: Predicting valence, arousal, and dominance. *The Quarterly Journal of Experimental Psychology*, *68*(8), 1584–1598. https://doi.org/10.1080/17470218.2014.941296

Recchia, G., & Louwerse, M. M. (2015b). Reproducing affective norms with lexical co-occurrence statistics: Predicting valence, arousal, and dominance. *The*

Quarterly Journal of Experimental Psychology, *68*(8), 1584–1598. https://doi.org/10.1080/17470218.2014.941296

Roelofs, A. (2004). Error biases in spoken word planning and monitoring by aphasic and nonaphasic speakers: Comment on rapp and goldrick (2000).

Roelofs, A. (2008). Attention, gaze shifting, and dual-task interference from phonological encoding in spoken word planning. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(6), 1580.

Roelofs, A. (2014). A dorsal-pathway account of aphasic language production: The weaver++/arc model. *Cortex*, *59*, 33–48.

Roelofs, A., & Piai, V. (2011). Attention demands of spoken word planning: A review. *Frontiers in Psychology*, *2*, 307.

Roest, S. A., Visser, T. A., & Zeelenberg, R. (2018). Dutch taboo norms. *Behavior Research Methods*, *50*(2), 630–641. https://doi.org/10.3758/s13428-017-0890-x

Rose, S. B., & Rahman, R. A. (2017). Semantic similarity promotes interference in the continuous naming paradigm: Behavioural and electrophysiological evidence. *Language, Cognition and Neuroscience*, *32*(1), 55–68. https://doi.org/10.1080/23273798.2016.1212081

Rosenzweig, M. R., & Postman, L. (1958). Frequency of Usage and the Perception of Words. *Science*, *127*(3293), 263–266.

Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, *8*(10), 627–633.

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, *2*, e55. https://doi.org/10.7717/peerj-cs.55

Sanborn, A. N., & Chater, N. (2016). Bayesian Brains without Probabilities. *Trends in Cognitive Sciences*, *20*(12), 883–893. https://doi.org/10.1016/j.tics.2016.10.003

Santilli, M., Vilas, M. G., Mikulan, E., Caro, M. M., Muñoz, E., Sedeño, L., Ibáñez, A., & García, A. M. (2019). Bilingual memory, to the extreme: Lexical processing in simultaneous interpreters. *Bilingualism: Language and Cognition*, *22*(2), 331–348. https://doi.org/10.1017/S1366728918000378

Scaltritti, M., Peressotti, F., & Navarrete, E. (2017). A joint investigation of semantic facilitation and semantic interference in continuous naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(5), 818–823. https://doi.org/10.1037/xlm0000354

Schauenburg, G., Ambrasat, J., Schröder, T., von Scheve, C., & Conrad, M. (2015). Emotional connotations of words related to authority and community. *Behavior Research Methods*, *47*(3), 720–735. https://doi.org/10.3758/s13428-014-0494-7

Schmidt, S., Scholl, P., Rensing, C., & Steinmetz, R. (2011). Cross-lingual recommendations in a resource-based learning scenario. In C. D. Kloos, D. Gillet, R. M. Crespo García, F. Wild, & M. Wolpers (Eds.), *Towards ubiquitous learning* (pp. 356–369). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-23985-4_28

Schweickert, R. (1980). Critical-path scheduling of mental processes in a dual task. *Science*, *209*(4457), 704–706.

Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., & Sereno, S. C. (2019). The Glasgow Norms: ratings of 5,500 words on nine scales. *Behavior Research Methods*, *51*(3), 1258–1270. https://doi.org/10.3758/s13428-018-1099-3

Seeber, K. G., & Kerzel, D. (2012). Cognitive load in simultaneous interpreting: Model meets data. *International Journal of Bilingualism*, *16*(2), 228–242. https://doi.org/10.1177/1367006911402982

Shao, Z., van Paridon, J., Poletiek, F., & Meyer, A. S. (2019). Effects of phrase and word frequencies in noun phrase production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(1), 147–165. https://doi.org/10.1037/xlm0000570

Sianipar, A., van Groenestijn, P., & Dijkstra, T. (2016). Affective meaning, concreteness, and subjective frequency norms for Indonesian words. *Frontiers in Psychology*, *7*, 1907. https://doi.org/10.3389/fpsyg.2016.01907

Soares, A. P., Comesaña, M., Pinheiro, A. P., Simões, A., & Frade, C. S. (2012). The adaptation of the Affective Norms for English Words (ANEW) for European Portuguese. *Behavior Research Methods*, *44*(1), 256–269. https://doi.org/10.3758/s13428-011-0131-7

Söderholm, C., Häyry, E., Laine, M., & Karrasch, M. (2013). Valence and arousal ratings for 420 finnish nouns by age and gender. *PloS one*, *8*(8), e72859. https://doi.org/10.1371/journal.pone.0072859

Speed, L. J., & Majid, A. (2017). Dutch modality exclusivity norms: Simulating perceptual modality in space. *Behavior Research Methods*, *49*(6), 2204–2218. https://doi.org/10.3758/s13428-017-0852-3

Stadthagen-Gonzalez, H., Imbault, C., Pérez Sánchez, M. A., & Brysbaert, M. (2017). Norms of valence and arousal for 14,031 Spanish words. *Behavior*

Research Methods, *49*(1), 111–123. https://doi.org/10.3758/s13428-015-0700-2

Stadthagen-González, H., Ferré, P., Pérez-Sánchez, M. A., Imbault, C., & Hinojosa, J. A. (2018). Norms for 10,491 Spanish words for five discrete emotions: Happiness, disgust, anger, fear, and sadness. *Behavior Research Methods*, *50*(5), 1943–1952. https://doi.org/10.3758/s13428-017-0962-y

Stavrakaki, S., Megari, K., Kosmidis, M. H., Apostolidou, M., & Takou, E. (2012). Working memory and verbal fluency in simultaneous interpreters. *Journal of Clinical and Experimental Neuropsychology*, *34*(6), 624–633.

Strijkers, K., Costa, A., & Thierry, G. (2010). Tracking Lexical Access in Speech Production: Electrophysiological Correlates of Word Frequency and Cognate Effects. *Cerebral Cortex*, *20*(4), 912–928. https://doi.org/10.1093/cercor/bhp153

Szumlanski, S., Gomez, F., & Sims, V. K. (2013). A new set of norms for semantic relatedness measures. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, *2*, 890–895.

Thompson, B., Roberts, S., & Lupyan, G. (2018). Quantifying semantic similarity across languages. *Proceedings of the 40th Annual Conference of the Cognitive Science Society (CogSci 2018)*.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288.

Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). Processing Advantages of Lexical Bundles: Evidence From Self-Paced Reading and Sentence Recall Tasks. *Language Learning*, *61*(2), 569–613. https://doi.org/10.1111/j.1467-9922.2010.00622.x

Turney, P. D., & Littman, M. L. (2002). Unsupervised learning of semantic orientation from a hundred-billion-word corpus.

Turney, P. D., & Littman, M. L. (2003). Measuring praise and criticism. *ACM Transactions on Information Systems*, *21*(4), 315–346. https://doi.org/10.1145/944012.944013

Tyler, L. K., Moss, H. E., Galpin, A., & Voice, J. K. (2002). Activating meaning in time: The role of imageability and form-class. *Language and Cognitive Processes*, *17*(5), 471–502.

Ueno, T., Saito, S., Rogers, T. T., & Ralph, M. A. L. (2011). Lichtheim 2: Synthesizing aphasia and the neural basis of language in a neurocomputational model of the dual dorsal-ventral language pathways. *Neuron*, *72*(2), 385–396.

Vankrunkelsven, H., Verheyen, S., De Deyne, S., & Storms, G. (2015). Predicting lexical norms using a word association corpus. *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, 2463–2468.

Venekoski, V., & Vankka, J. (2017). Finnish resources for evaluating language model semantics. *Proceedings of the Nordic Conference on Computational Linguistics*.

Vergallito, A., Petilli, M. A., & Marelli, M. (2020). Perceptual modality norms for 1,121 Italian words: A comparison with concreteness and imageability scores and an analysis of their impact in word processing tasks. *Behavior Research Methods*, 1–18. https://doi.org/10.3758/s13428-019-01337-8

Verheyen, S., De Deyne, S., Linsen, S., & Storms, G. (2019). Lexicosemantic, affective, and distributional norms for 1,000 Dutch adjectives. *Behavior Research Methods*, 1–14. https://doi.org/10.3758/s13428-019-01303-4

Walker, G. M., & Hickok, G. (2016). Bridging computational approaches to speech production: The semanticlexicalauditorymotor model (slam). *Psychonomic Bulletin & Review*, *23*(2), 339–352.

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

Westbury, C. F., Shaoul, C., Hollis, G., Smithson, L., Briesemeister, B. B., Hofmann, M. J., & Jacobs, A. M. (2013). Now you see it, now you don't: On emotion, context, and the algorithmic prediction of human imageability judgments. *Frontiers in Psychology*, *4*, 991. https://doi.org/10.3389/fpsyg.2013.00991

Woumans, E., Ceuleers, E., der Linden, L. V., Szmalec, A., & Duyck, W. (2015). Verbal and nonverbal cognitive control in bilinguals and interpreters. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *41*(5), 1579–1586.

Yang, D., & Powers, D. M. (2006). *Verb similarity on the taxonomy of WordNet*. Masaryk University.

Yap, M. J., Liow, S. J. R., Jalil, S. B., & Faizal, S. S. B. (2010). The Malay Lexicon Project: A database of lexical statistics for 9,592 words. *Behavior Research Methods*, *42*(4), 992–1003. https://doi.org/10.3758/BRM.42.4.992

Yudes, C., Macizo, P., & Bajo, T. (2011). The influence of expertise in simultaneous interpreting on non-verbal executive processes. *Frontiers in psychology*, *2*, 309.

Zesch, T., & Gurevych, I. (2006). Automatically creating datasets for measures of semantic relatedness. *Proceedings of the Workshop on Linguistic Distances*.

Zwitserlood, P. (1989). The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition, 32*(1), 25–64.

Nederlandse samenvatting

Op grote congressen en videobeelden van EU- en VN-vergaderingen zie je ze wel eens: simultaantolken. In plaats van eerst te luisteren en dan te vertalen (zoals andere tolken doen), vertalen ze simultaan, terwijl ze luisteren. Dit is geen gemakkelijke taak, in de eerste plaats omdat het niet makkelijk is iemand anders te verstaan terwijl je zelf spreekt, maar ook omdat het vereist dat je je aandacht verdeelt over twee taken (spreken en luisteren) in twee verschillende talen. Omdat simultaantolken zo moeilijk is, is het een interessante taak voor taalpsychologen. Als we kunnen achterhalen welke spreek- en luisterprocessen tijdens het tolken bemoeilijkt worden omdat ze tegelijk uitgevoerd moeten worden—en welke processen niet lijden onder het tolken—dan kan dat nieuwe inzichten verschaffen in hoe deze spreek- en luisterprocessen onder normale omstandigheden werken.

In Hoofdstuk 2 en 3 stel ik, met het oog op het beter begrijpen van deze spreek- en luisterprocessen, een tolkmodel voor dat gebaseerd is op taalpsychologische kennis uit eerder onderzoek. Het model voorspelt wanneer een tolk (gemiddeld genomen) fouten maakt, op basis van het spreektempo van de tekst die getolkt moet worden. Een centrale aanname in het model is dat het te complex is om woorden te begrijpen in de ene taal en tegelijkertijd de juiste woorden te bedenken in de andere taal en dat de hersenen deze processen daarom afwisselen. Dit model test ik vervolgens op opnames van proefpersonen die een eenvoudige tolktaak uitvoeren. In Hoofdstuk 2 tolken de proefpersonen eenvoudige, korte verhaaltjes uit kinderboeken. Na afstellen van de parameters van het model blijkt dat het de gemiddelde foutpercentages van proefpersonen correct kan nabootsen, mits we aannemen dat het een kort moment (ongeveer 50 milliseconden) duurt om te schakelen tussen het begrijpen van woorden in de ene taal en het bedenken van de juiste woorden in de andere taal. In Hoofdstuk 3 onderzoek ik of dit ook geldt voor het tolken van lijstjes van losse woorden, wat proefpersonen als moeilijker ervaren omdat ze geen gebruik kunnen maken van semantische en syntactische context om woorden beter te begrijpen of vertalen. Het model blijkt voor zulke lijstjes niet te werken, wat verrassend is omdat het model alleen

naar spreeksnelheid kijkt en de afwezigheid van context dus geen invloed zou moeten hebben. Anders bezien is het misschien juist verrassend dat het model goed lijkt te werken voor de verhaaltjes uit Hoofdstuk 2, ondanks dat het geen rekening houdt met de semantische en syntactische context die de verhaaltjes verschaffen.

Hoe goed en snel een woord getolkt kan worden is dus deels afhankelijk van de context, maar welke factoren dragen daar precies aan bij? In hoeverre is het huidige woord voorspelbaar op basis van het voorgaande woord? Is het semantisch en syntactisch inpasbaar in de context? Het model uit Hoofdstuk 2 en 3 geeft op deze vragen helaas geen antwoord. In Hoofdstuk 4 maak ik daarom gebruik van recent ontwikkelde statistische methodes die het mogelijk maken om de invloed van complexe factoren als semantiek en syntax op de spreeksnelheid van proefpersonen die tolken te modelleren. De resultaten van deze statistische analyse laten zien dat veel van de factoren die van invloed zijn op normaal spreken en luisteren ook invloed hebben op tolken, maar dat de mate waarin zij van invloed zijn afhangt van de specifieke taak die proefpersonen uitvoeren. Het maken van een computermodel zoals in Hoofdstuk 2 en 3, dat de spreek- en luisterprocessen tijdens tolken simuleert, kan dus eigenlijk niet zonder ook complexe semantische en syntactische factoren te modelleren. Het kerndeel van de dissertatie komt hiermee ten einde, maar in de resterende twee hoofdstukken presenteer ik werk dat fungeerde als bouwstenen voor het onderzoek in de eerdere hoofdstukken.

Om bijvoorbeeld, zoals ik doe in Hoofdstuk 4, het effect van context op de spreeksnelheid van tolken inzichtelijk te maken, moet ik die context eerst numeriek kunnen weergeven. In Hoofdstuk 6 gebruik ik een techniek uit het Machine Learning-veld om numerieke voorstellingen van de semantiek van losse woorden te verkrijgen. Het basisprincipe is eenvoudig: Een zelflerend (neural network) model leest ondertitels en probeert op basis van ieder woord de omliggende woorden te voorspellen. Als het model fout gokt dan leert het en stelt het zijn interne numerieke voorstellingen voor de voorspelde woorden bij. Door dit proces vele malen te herhalen voor een enorm archief van ondertitels leert het steeds beter te voorspellen welke woorden in een bepaalde context kunnen voorkomen. De interne numerieke voorstellingen die het model leert kunnen we na afloop van het leerproces uit het model halen en zelf gebruiken om te voorspellen welke woorden semantische gelijkenis vertonen.

We kunnen ook met een nóg eenvoudiger methode de voorspelbaarheid van een woord op basis van de voorgaande context vaststellen: Door simpelweg

te tellen hoe vaak een woord op die voorgaande context volgt in een enorm corpus van voorbeeldzinnen (we kunnen hiervoor wederom het archief van ondertitels gebruiken). Met deze methode kunnen we de kans dat een woord volgt op de voorgaande context berekenen, maar gebruikelijker in taalkundig onderzoek is om het logaritme van de kans te gebruiken. Deze index wordt ook wel de *surprisal* (verrassing) genoemd. Deze waarde kan ook andersom berekend worden: Hoe verrassend is de voorgaande context, wanneer we het huidige woord als uitgangspunt nemen? Beide waarden worden regelmatig gebruikt in taalkundig onderzoek, maar in Hoofdstuk 5 beschrijf ik hoe dit een probleem kan vormen voor het interpreteren van onderzoeksresultaten. Er bestaan wiskundige verbanden tussen de verschillende manieren om verrassingswaarden te berekenen voor losse woorden en stukken tekst. Door deze verbanden kan een statistisch model dat al deze factoren probeert te vergelijken paradoxaal niet goed onderscheid maken tussen de verschillende factoren, een verschijnsel dat we multicollineariteit noemen. Er zijn verschillende methoden om met multicollineariteit om te gaan, maar aan deze methoden kleven ook weer nadelen. Ik concludeer dat het vaak beter is om zorgvuldig te overwegen of er a priori factoren geselecteerd kunnen worden die op theoretische gronden relevant zijn en andere factoren buiten beschouwing te laten.

Acknowledgements

I owe a word of thanks to a number of people in- and outside the MPI. Some were essential to the research reported in this thesis, others kept me sane while I was doing said research. All were instrumental in getting me through the past five years. Thank you.

To my room 362 officemates: Joe, for heckling me at every turn (for my own good, of course). You started your PhD after me and finished it before me, as you had predicted all along. Eirini, for helping me drink all those gallons of coffee we made. You showed me that even people more conscientious than me still worry constantly about the work they are producing, which was weirdly comforting.

To my labmates and friends, you made the MPI a kinder place.

To the MPI staff: All the operations, library, and technical staff that keep the MPI going day after day, but especially Tobias, for keeping the MPIs compute cluster and file servers going despite our heavy use (and occasional misuse). This dissertation could not have been completed without your patient support. Annelies and the Psychology of Language research assistants: This project required lots of annotation which took many hours of research assistant time. It can't have been much fun, but I hope it didn't leave any lasting scars. Kevin, for organizing and helping to organize the IMPRS curriculum and IMPRS conference. It must be a thankless job at times, but I can't imagine what the IMPRS would look like without you.

To my promotors: Ardi, for your sharp eye and encouragement not to overcomplicate things. Antje, for giving me this opportunity, and for your patience while I struggled to finish this dissertation.

To my collaborators: Fenna, my undergraduate and Masters thesis advisor, for involving me in my first real research project and for suggesting I apply for an internship at the MPI. Without you I would have never found my way here. Zeshu, for supervising my internship project at the MPI, paving the way to the start of this PhD project. Bill, for your enthusiasm in engaging with any open question, including mine. Our collaboration set me on the path I am on now. Falk, for involving me in your globe-spanning research project, and for being generous with

credit and insight. Guillermo, for your friendly encouragement, without which I would still be stuck halfway up a volcano somewhere out in the Atlantic Ocean.

To Markus, for involving me in your PhD research when I was an intern. Your breadth of knowledge and appetite for new and interesting projects helped motivate me in times when my thesis research was progressing slower than I would have liked.

To Phillip, for all your help and guidance (and coffee beans). There were many times I surely would have abandoned this dissertation if not for your encouragement and commiseration.

To my family: My parents, for supporting me through roughly 25 years of schooling. I know I meandered a bit but I think I'm done now. Suus, for sharing my capacity for worrying, and for understanding it has nothing to do with how well things are going on paper.

To Jack and Liz, for putting up with me staying at your house for weeks at a time. Occasionally putting physical distance between myself and the MPI helped keep me sane.

To Scott, for our many long discussions about things that do not relate to this dissertation in any way whatsoever.

To Roemer and Daniel, for being my oldest friends.

To Ashley, for your love and support, through all of *this*. I promise this is the last time I get a PhD.

Curriculum Vitae

Jeroen van Paridon (Voorburg, the Netherlands, 1990) graduated primary school in 2001 without ever having learned to write cursive. He later obtained a Bachelor's in Psychology from Leiden University in 2012, followed by a Master's in Psychology from that same university in 2015. He began his PhD research at the Max Planck Institute for Psycholinguistics in Nijmegen later that year. Some of his work there ended up in this dissertation; you can find the rest of it online at jvparidon.io/publications. At the time of printing, Jeroen is working as a postdoctoral research scholar at the University of Wisconsin in Madison.

Publications

Montero-Melis, G., Isaksson, P., **Van Paridon, J.,** & Ostarek, M. (2020). Does using a foreign language reduce mental imagery? *Cognition, 196*, 104134. https://doi.org/10.1016/j.cognition.2019.104134

Ostarek, M., **Van Paridon, J.,** & Montero-Melis, G. (2020). Sighted peoples language is not helpful for blind individuals acquisition of typical animal colors. *Proceedings of the National Academy of Sciences, 116*(44), 21972–21973. https://doi.org/10.1073/pnas.1912302116

Shao, Z., **Van Paridon, J.,** Poletiek, F., & Meyer, A. S. (2019). Effects of phrase and word frequencies in noun phrase production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*(1), 147-165. https://doi.org/10.1037/xlm0000570

Van Paridon, J., Ostarek, M., Arunkumar, M., & Huettig, F. (2020). Does neuronal recycling result in destructive competition? The influence of learning to read on the recognition of faces. *Psychological Science*. https://doi.org/10.1177%2F0956797620971652

Van Paridon, J. & Thompson, B. (2020). subs2vec: Word embeddings from subtitles in 55 languages. *Behavior Research Methods*. https://doi.org/10.3758/s13428-020-01406-3

Van Paridon, J., Roelofs, A., & Meyer, A. S. (2019). A lexical bottleneck in shadowing and translating of narratives. *Language, Cognition and Neuroscience, 34*(6), 803–812. https://doi.org/10.1080/23273798.2019.1591470

Warren, C. M., Tona, K. D., Ouwerkerk, L., **Van Paridon, J.,** Poletiek, F., Van Steenbergen, H., Bosch, J. A., & Nieuwenhuis, S. (2019). The neuromodulatory and hormonal effects of transcutaneous vagus nerve stimulation as evidenced by salivary alpha amylase, salivary cortisol, pupil diameter, and the P3 event-related potential. *Brain Stimulation, 12*(3), 635–642. https://doi.org/10.1016/j.brs.2018.12.224

MPI Series in Psycholinguistics

1. The electrophysiology of speaking: Investigations on the time course of semantic, syntactic, and phonological processing. *Miranda I. van Turennout*
2. The role of the syllable in speech production: Evidence from lexical statistics, metalinguistics, masked priming, and electromagnetic midsagittal articulography. *Niels O. Schiller*
3. Lexical access in the production of ellipsis and pronouns. *Bernadette M. Schmitt*
4. The open-/closed class distinction in spoken-word recognition. *Alette Petra Haveman*
5. The acquisition of phonetic categories in young infants: A self-organising artificial neural network approach. *Kay Behnke*
6. Gesture and speech production. *Jan-Peter de Ruiter*
7. Comparative intonational phonology: English and German. *Esther Grabe*
8. Finiteness in adult and child German. *Ingeborg Lasser*
9. Language input for word discovery. *Joost van de Weijer*
10. Inherent complement verbs revisited: Towards an understanding of argument structure in Ewe. *James Essegbey*
11. Producing past and plural inflections. *Dirk J. Janssen*
12. Valence and transitivity in Saliba: An Oceanic language of Papua New Guinea. *Anna Margetts*
13. From speech to words. *Arie H. van der Lugt*
14. Simple and complex verbs in Jaminjung: A study of event categorisation in an Australian language. *Eva Schultze-Berndt*
15. Interpreting indefinites: An experimental study of children's language comprehension. *Irene Krämer*

16. Language-specific listening: The case of phonetic sequences. *Andrea Christine Weber*
17. Moving eyes and naming objects. *Femke Frederike van der Meulen*
18. Analogy in morphology: The selection of linking elements in Dutch compounds. *Andrea Krott*
19. Morphology in speech comprehension. *Kerstin Mauth*
20. Morphological families in the mental lexicon. *Nivja Helena de Jong*
21. Fixed expressions and the production of idioms. *Simone Annegret Sprenger*
22. The grammatical coding of postural semantics in Goemai (a West Chadic language of Nigeria. *Birgit Hellwig*
23. Paradigmatic structures in morphological processing: Computational and cross-linguistic experimental studies. *Fermín Moscoso del Prado Martín*
24. Contextual influences on spoken-word processing: An electrophysiological approach. *Danielle van den Brink*
25. Perceptual relevance of prevoicing in Dutch. *Petra Martine van Alphen*
26. Syllables in speech production: Effects of syllable preparation and syllable frequency. *Joana Cholin*
27. Producing complex spoken numerals for time and space. *Marjolein Henriëtte Wilhelmina Meeuwissen*
28. Morphology in auditory lexical processing: Sensitivity to fine phonetic detail and insensitivity to suffix reduction. *Rachèl Jenny Judith Karin Kemps*
29. At the same time: The expression of simultaneity in learner varieties. *Barbara Schmiedtová*
30. A grammar of Jalonke argument structure. *Friederike Lüpke*
31. Agrammatic comprehension: An electrophysiological approach. *Marijtje Elizabeth Debora Wassenaar*
32. The structure and use of shape-based noun classes in Miraña (North West Amazon). *Frank Seifart*
33. Prosodically-conditioned detail in the recognition of spoken words. *Anne Pier Salverda*
34. Phonetic and lexical processing in a second language. *Mirjam Elisabeth Broersma*
35. Retrieving semantic and syntactic word properties: ERP studies on the time course in language comprehension. *Oliver Müller*

36. Lexically-guided perceptual learning in speech processing. *Frank Eisner*

37. Sensitivity to detailed acoustic information in word recognition. *Keren Batya Shatzman*

38. The relationship between spoken word production and comprehension. *Rebecca Özdemir*

39. Disfluency: Interrupting speech and gesture. *Mandana Seyfeddinipur*

40. The acquisition of phonological structure: Distinguishing contrastive from non-constrative variation. *Christiane Dietrich*

41. Cognitive cladistics and the relativity of spatial cognition. *Daniel Haun*

42. The acquisition of auditory categories. *Martijn Bastiaan Goudbeek*

43. Affix reduction in spoken Dutch: Probabilistic effects in production and perception. *Mark Plumaekers*

44. Continuous-speech segmentation at the beginning of language acquisition: Electrophysiological evidence. *Valesca Madalla Kooijman*

45. Space and iconicity in German sign language (DGS). *Pamela M. Perniss*

46. On the production of morphologically complex words with special attention to effects of frequency. *Heidrun Bien*

47. Crosslinguistic influence in first and second languages: Convergence in speech and gesture. *Amanda Brown*

48. The acquisition of verb compounding in Mandarin Chinese. *Jidong Chen*

49. Phoneme inventories and patterns of speech sound perception. *Anita Eva Wagner*

50. Lexical processing of morphologically complex words: An information-theoretical perspective. *Victor Kuperman*

51. A grammar of Savosavo: A Papuan language of the Solomon Islands. *Claudia Ursula Wegener*

52. Prosodic structure in speech production and perception. *Claudia Kuzla*

53. The acquisition of finiteness by Turkish learners of German and Turkish learners of French: Investigating knowledge of forms and functions in production and comprehension. *Sarah Schimke*

54. Studies on intonation and information structure in child and adult German. *Laura de Ruiter*

55. Processing the fine temporal structure of spoken words. *Eva Reinisch*

56. Semantics and (ir)regular inflection in morphological processing. *Wieke Tabak*

57. Processing strongly reduced forms in casual speech. *Susanne Brouwer*

58. Ambiguous pronoun resolution in L1 and L2 German and Dutch. *Miriam Ellert*

59. Lexical interactions in non-native speech comprehension: Evidence from electroencephalography, eye-tracking, and functional magnetic resonance imaging. *Ian FitzPatrick*

60. Processing casual speech in native and non-native language. *Annelie Tuinman*

61. Split intransitivity in Rotokas, a Papuan language of Bougainville. *Stuart Payton Robinson*

62. Evidentiality and intersubjectivity in Yurakaré: An interactional account. *Sonja Gipper*

63. The influence of information structure on language comprehension: A neurocognitive perspective. *Lin Wang*

64. The meaning and use of ideophones in Siwu. *Mark Dingemanse*

65. The role of acoustic detail and context in the comprehension of reduced pronunciation variants. *Marco van de Ven*

66. Speech reduction in spontaneous French and Spanish. *Francisco Torreira*

67. The relevance of early word recognition: Insights from the infant brain. *Caroline Mary Magteld Junge*

68. Adjusting to different speakers: Extrinsic normalization in vowel perception. *Matthias Johannes Sjerps*

69. Structuring language: Contributions to the neurocognition of syntax. *Katrien Rachel Segaert*

70. Infants' appreciation of others' mental states in prelinguistic communication: A second person approach to mindreading. *Birgit Knudsen*

71. Gaze behavior in face-to-face interaction. *Federico Rossano*

72. Sign-spatiality in Kata Kolok: How a village sign language of Bali inscribes its signing space. *Connie de Vos*

73. Who is talking? Behavioural and neural evidence for norm-based coding in voice identity learning. *Attila Andics*

74. Lexical processing of foreign-accented speech: Rapid and flexible adaptation. *Marijt Witteman*

75. The use of deictic versus representational gestures in infancy. *Daniel Puccini*

76. Territories of knowledge in Japanese conversation. *Kaoru Hayano*

77. Family and neighbourhood relations in the mental lexicon: A cross-language perspective. *Kimberley Mulder*

78. Contributions of executive control to individual differences in word production. *Zeshu Shao*

79. Hearing speech and seeing speech: Perceptual adjustments in auditory-visual processing. *Patrick van der Zande*

80. High pitches and thick voices: The role of language in space-pitch associations. *Sarah Dolscheid*

81. Seeing what's next: Processing and anticipating language referring to objects. *Joost Rommers*

82. Mental representation and processing of reduced words in casual speech. *Iris Hanique*

83. The many ways listeners adapt to reductions in casual speech. *Katja Pöllmann*

84. Contrasting opposite polarity in Germanic and Romance languages: Verum Focus and affirmative particles in native speakers and advanced L2 learners. *Giuseppina Turco*

85. Morphological processing in younger and older people: Evidence for flexible dual-route access. *Jana Reifegerste*

86. Semantic and syntactic constraints on the production of subject-verb agreement. *Alma Veenstra*

87. The acquisition of morphophonological alternations across languages. *Helen Buckler*

88. The evolutionary dynamics of motion event encoding. *Annemarie Verkerk*

89. Rediscovering a forgotten language. *Jiyoun Choi*

90. The road to native listening: Language-general perception, language-specific input. *Sho Tsuji*

91. Infants'understanding of communication as participants and observers. *Gudmundur Bjarki Thorgrímsson*

92. Information structure in Avatime. *Saskia van Putten*

93. Switch reference in Whitesands. *Jeremy Hammond*

94. Machine learning for gesture recognition from videos. *Binyam Gebrekidan Gebre*

95. Acquisition of spatial language by signing and speaking children: A comparison of Turkish sign language (TID) and Turkish. *Beyza Sumer*

96. An ear for pitch: On the effects of experience and aptitude in processing pitch in language and music. *Salomi Savvatia Asaridou*

97. lncrementality and Flexibility in Sentence Production. *Maartje van de Velde*

98. Social learning dynamics in chimpanzees: Reflections on (nonhuman) animal culture. *Edwin van Leeuwen*

99. The request system in Italian interaction. *Giovanni Rossi*

100. Timing turns in conversation: A temporal preparation account. *Lilla Magyari*

101. Assessing birth language memory in young adoptees. *Wencui Zhou*

102. A social and neurobiological approach to pointing in speech and gesture. *David Peeters*

103. Investigating the genetic basis of reading and language skills. *Alessandro Gialluisi*

104. Conversation electrified: The electrophysiology of spoken speech act recognition. *Rósa Signý Gisladottir*

105. Modelling multimodal language processing. *Alastair Charles Smith*

106. Predicting language in different contexts: The nature and limits of mechanisms in anticipatory language processing. *Florian Hintz*

107. Situational variation in non-native communication *Huib Kouwenhoven*

108. Sustained attention in language production *Suzanne Jongman*

109. Acoustic reduction in spoken-word processing: Distributional, syntactic, morphosyntactic, and orthographic effects *Malte Viebahn*

110. Nativeness, dominance, and the flexibility of listening to spoken language *Laurence Bruggeman*

111. Semantic specificity of perception verbs in Maniq *Ewelina Wnuk*

112. On the identification of FOXP2 gene enhancers and their role in brain development *Martin Becker*

113. Events in language and thought: The case of serial verb constructions in Avatime *Rebecca Defina*

114. Deciphering common and rare genetic effects on reading ability *Amaia Carrión Castillo*

115. Music and language comprehension in the brain *Richard Kunert*

116. Comprehending Comprehension: Insights from neuronal oscillations on the neuronal basis of language *Nietzsche H.L. Lam*

117. The biology of variation in anatomical brain asymmetries *Tulio Guadalupe*

118. Language processing in a conversation context *Lotte Schoot*

119. Achieving mutual understanding in Argentine Sign Language *Elizabeth Manrique*

120. Talking Sense: the behavioural and neural correlates of sound symbolism *Gwilym Lockwood*

121. Getting under your skin: The role of perspective and simulation of experience in narrative comprehension *Franziska Hartung*

122. Sensorimotor Experience in Speech Perception *Will Schuerman*

123. Explorations of beta-band neural oscillations during language comprehension: Sentence processing and beyond *Ashley Lewis*

124. Influences on the magnitude of syntactic priming *Evelien Heyselaar*

125. Lapse organization in interaction *Elliott Hoey*

126. The processing of reduced word pronunciation variants by natives and foreign language learners: Evidence from French casual speech *Sophie Brand*

127. The Neighbors Will Tell You What To Expect: Effects of Aging and Predictability on Language Processing *Cornelia Moers*

128. The role of voice and word order in incremental sentence processing. Studies on sentence production and comprehension in Tagalog and German *Sebastian Sauppe*

129. Learning from the (Un)Expected: Age and Individual Differences in Statistical Learning and Perceptual Learning in Speech *Thordis Neger*

130. Mental representations of Dutch regular morphologically complex neologisms *Laura de Vaan*

131. Speech production, perception, and input of simultaneous bilingual preschoolers: Evidence from voice onset time *Antje Stoehr*

132. A holistic approach to understanding pre-history *Vishnupriya Kolipakam*

133. Characterization of transcription factors in monogenic disorders of speech and language *Sara Busquets Estruch*

134. Indirect request comprehension in different contexts *Johanne Tromp*

135. Envisioning Language - An Exploration of Perceptual Processes in Language Comprehension *Markus Ostarek*

136. Listening for the WHAT and the HOW: Older adults processing of semantic and affective information in speech *Juliane Kirsch*

137. Let the agents do the talking: On the influence of vocal tract anatomy on speech during ontogeny and glossogeny *Rick Janssen*

138. Age and hearing loss effects on speech processing *Xaver Koch*

139. Vocabulary knowledge and learning: Individual differences in adult native speakers *Nina Mainz*

140. The face in face-to-face communication: Signals of understanding and non-understanding *Paul Hömke*

141. Person reference and interaction in Umpila/Kuuku Yau narrative *Clair Hill*

142. Beyond the language given: The neurobiological infrastructure for pragmatic inferencing *Jana Banáková*

143. From Kawapanan to Shawi: Topics in language variation and change *Luis Miguel Rojas Berscia*

144. On the oscillatory dynamics underlying speech-gesture integration in clear and adverse listening conditions *Linda Drijvers*

145. Linguistic dual-tasking: Understanding temporal overlap between production and comprehension *Amie Fairs*

146. The role of exemplars in speech comprehension *Annika Nijveld*

147. A network of interacting proteins disrupted in language-related disorders *Elliot Sollis*

148. Fast speech can sound slow: Effects of contextual speech rate on word recognition *Merel Maslowski*

149. Reasons for every-day activities *Julija Baranova*

150. Speech planning in dialogue - Psycholinguistic studies of the timing of turn taking *Mathias Barthel*

151. Exploring social biases in language processing *Sara Iacozza*

152. The role of neural feedback in language unification: How awareness affects combinatorial processing *Valeria Mongelli*

153. Vocal learning in the pale spear-nosed bat, *Phyllostomus discolor Ella Lattenkamp*

154. The effect of language contact on speech and gesture: The case of Turkish-Dutch bilinguals in the Netherlands *Elif Zeynep Azar*

155. Language and society: How social pressures shape grammatical structure *Limor Raviv*

156. The moment in between: Planning speech while listening *Svetlana-Lito Gerakaki*

157. How speaking fast is like running: Modelling control of speaking rate *Joe Rodd*

158. The power of context: How linguistic contextual information shapes brain dynamics during sentence processing *René Terporten*

159. Neurobiological models of sentence processing *Marvin Uhlmann*

160. Individual differences in syntactic knowledge and processing: The role of literacy experience *Saoradh Favier*

161. Memory for speaking and listening *Eirini Zormpa*

162. Masculine generic pronouns: Investigating the processing of an unintended gender cue *Theresa Redl*

163. Properties, structures and operations: Studies on language processing in the brain using computational linguistics and naturalistic stimuli *Alessandro Lopopolo*

164. Investigating spoken language comprehension as perceptual inference *Greta Kaufeld*

165. What was that Spanish word again? Investigations into the cognitive mechanisms underlying foreign language attrition *Anne Mickan*

166. A tale of two modalities: How modality shapes language production and visual attention *Francie Manhardt*

167. Why do we change how we speak? Multivariate genetic analyses of language and related traits across development and disorder *Ellen Verhoef*

168. Variation in form and meaning across the Japonic language family with a focus on the Ryukyuan languages *John Huisman*

169. Bilingual sentence production and code-switching: Neural network simulations *Chara Tsoukala*

170. Effects of aging and cognitive abilities on multimodal language production and comprehension in context *Louise Schubotz*

171. Speaking while listening: Language processing in speech shadowing and translation *Jeroen van Paridon*