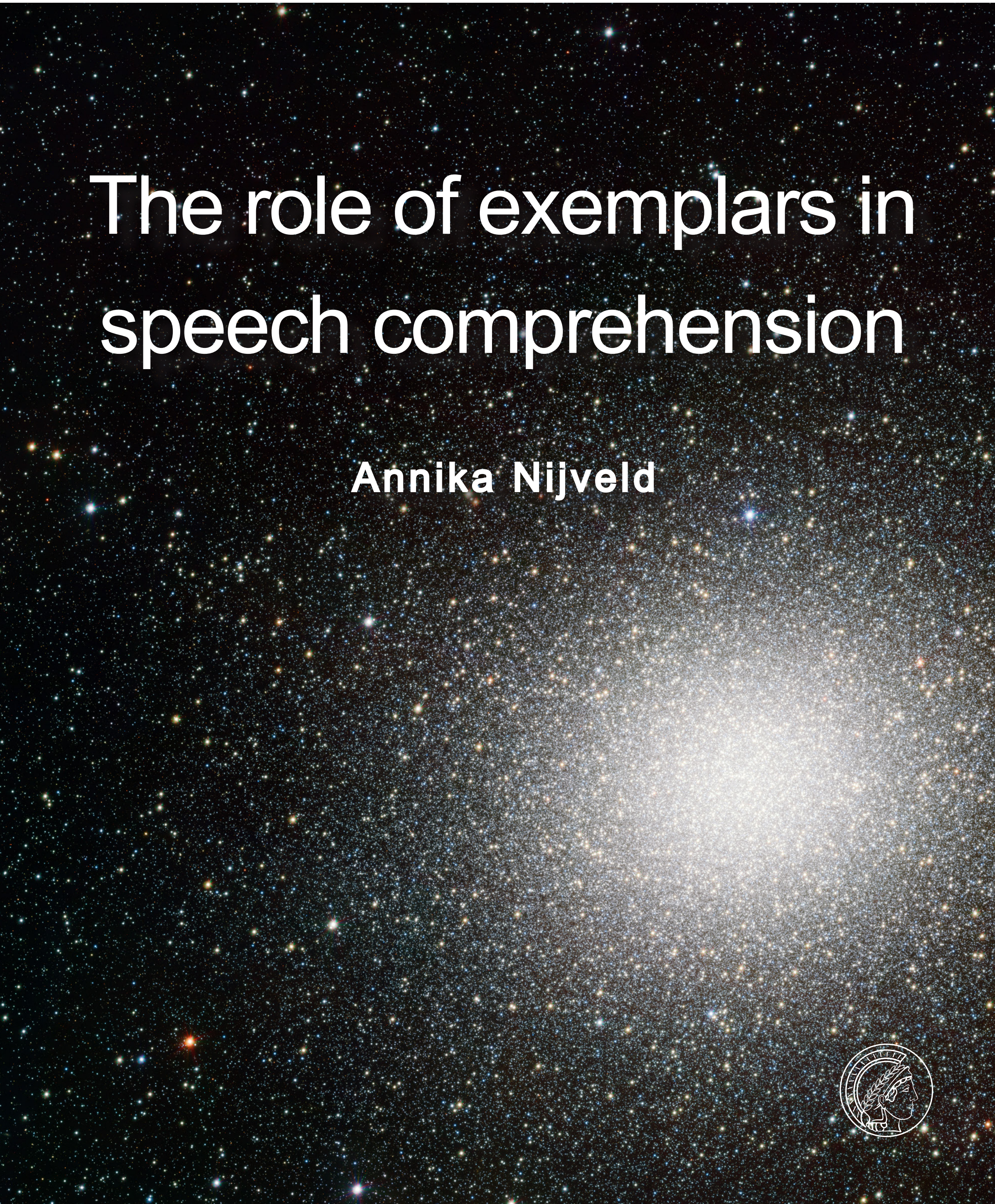
MAX PLANCK INSTITUTE
FOR PSYCHOLINGUISTICS

WWW.MPI.NL



The role of exemplars in speech comprehension

Annika Nijveld

The role of exemplars in speech comprehension

The research reported in this thesis was supported by a VICI grant from the Netherlands Organisation for Scientific Research (NWO; grant number 277-70-010) awarded to Mirjam Ernestus.

Cover image displays the largest globular star cluster in the sky, Omega Centauri. Credit: ESO/INAF-VST/OmegaCAM. Acknowledgement: A. Grado, L. Limatola/INAF-Capodimonte Observatory.

ISBN: 978-94-92910-03-5
Printed and bound by Ipskamp Printing, Enschede

The role of exemplars in speech comprehension

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 decanen
in het openbaar te verdedigen op maandag 1 juli 2019
om 10.30 uur precies

door

Annika Dorinthe Nijveld

geboren op 2 november 1987
te Voorburg

Promotor:

Prof. dr. M.T.C. Ernestus

Copromotor:

Dr. L.F.M. ten Bosch

Manuscriptcommissie:

Prof. dr. R.W.N.M. van Hout	(voorzitter)
Dr. M.E. Broersma	
Prof. dr. A. Chen	(Universiteit Utrecht)
Prof. dr. J.M. McQueen	
Dr. A. Schweitzer	(Universität Stuttgart, Duitsland)

Contents

Introduction

Chapter 1

A long-standing question in research on speech comprehension is how the pronunciation of words is represented in listeners' minds. Traditional 'abstractionist' approaches assume that a typical word has a single lexical representation, which consists of a sequence of abstract units such as phonemes. Situation and speaker-specific information of tokens is not retained in such abstract lexical representations. Other, 'exemplar-based' approaches, assume that a word's pronunciation is stored as many individual tokens, which together form a cloud for that word. These tokens are represented in full phonetic detail, and thus retain situation- and speaker-specific information. In a final, hybrid class of models of speech comprehension, elements of abstractionist and exemplar-based approaches are mixed, and these models posit that abstract representations and exemplars coexist in listeners' minds. Although most researchers in the field now agree that listeners retain in memory both abstract and detailed information of a word's pronunciation, there are still many unanswered questions. In this dissertation, we further investigate the nature of exemplars and their role in speech comprehension, and what this tells us about the speech comprehension process.

Abstractionist models

The traditional abstractionist view of speech comprehension has its roots in structuralism (e.g., Trubetzkoy, 1939), and was later adopted by generative linguistics. This view dominated psycholinguistics and speech comprehension research in the 1970s and 1980s. It posits that words are represented in the mental lexicon as economically as possible: word representations only code linguistically contrastive information (i.e., information needed to distinguish one word from another), while any additional information present in the speech signal (i.e., information associated with the situational context and the speaker of a specific token) is discarded. In the words of Halle (1985, p. 101):

> "When we learn a new word, we practically never remember most of the salient acoustic properties that must have been present in the signal that struck our ears. For example, we do not remember the voice quality,

> speed of utterance, and other properties directly linked to the unique circumstances surrounding every utterance."

Abstractionist models assume that during speech processing, all extra-linguistic features of the speech signal are thus stripped away, or normalized, from words' linguistic content. The normalization process has been used to account for how listeners can recognize words spoken by many different unknown talkers and speech styles without substantial effort. Well-known abstractionist theoretical models of speech comprehension include the Cohort model (Marslen-Wilson & Welsh, 1978; Marslen-Wilson, 1987), TRACE (McClelland & Elman, 1986), and Shortlist (Norris, 1994); see McQueen (2005) for a more detailed discussion of several models.

Exemplar models

From the 1990s, the abstractionist view was challenged by exemplar (or episodic) approaches to speech comprehension (e.g., Goldinger, 1998; Johnson, 1997; Pierrehumbert, 2001). Exemplar theory was first introduced in psychology as a model of perception and categorization (Hintzman, 1986; Nosofsky, 1986). The central assumption of exemplar-based models is that perceivers encode richly detailed instances of experienced events (i.e., exemplars) into memory, which retain situation-specific information. Each category is represented by a cloud of experienced tokens of that category, and exemplars are organized in a cognitive map such that highly similar exemplars are close to each other, while dissimilar exemplars are far apart. Individual exemplars may cluster together in more than one category, such as the word and the talker in the case of speech comprehension. When a new token is encountered, it is classified according to its similarity (i.e., its distance) to stored exemplars. In exemplar-based models of speech comprehension, extra-linguistic information such as about the identity of the speaker is thus not stripped away: it is assumed to be retained in the mental lexicon.

The evidence in favour of the storage of spoken words as clouds of exemplars comes from a range of experimental studies. For instance, in shadowing experiments, Goldinger (1998) observed that listeners mimic the acoustic properties of the speech signal when they have to reproduce spoken input. This finding implies that listeners must have stored these extra-linguistic properties of the speech signal. Other studies carried out long-term identity priming experiments, in which words were repeated. How well the first ('prime') and second ('target') tokens of a word matched acoustically was varied as stimulus condition: Primes and targets were produced by the same or different speaker, had the same or a different realization of a certain segment, or were produced at the same or a different speech rate (e.g., Bradlow, Nygaard, & Pisoni, 1999; Craik & Kirsner, 1974; McLennan, Luce, & Charles-Luce, 2003; Strori, Zaar, Cooke, & Mattys, 2018). If the prime creates an exemplar, priming (i.e., faster and/or more accurate responses to the targets) should be larger when prime and target tokens are highly similar. In contrast, when both tokens are recognized via the same abstract representation, priming should be equivalent for similar and dissimilar primes and targets. Many of these experiments observed enhanced priming when prime and target were acoustically similar ('exemplar effects'). These findings again indicate that listeners, at least temporarily, retain extra-linguistic information in memory.

Hybrid models

A third class of models emerged at the beginning of this century, and consists of so-called "hybrid models" (e.g., Goldinger, 2007; McLennan et al., 2003; Pierrehumbert, 2002). These models occupy a middle ground between abstractionist and exemplar-based models, and assume that a word's pronunciation is stored in the mental lexicon both in the form of an abstract representation and as exemplars.

Hybrid models originate from researchers' wish to incorporate the advantages of exemplars and abstract representations into a single model, as researchers have argued that neither purely abstractionist models nor purely exemplar-based models can account for all the findings in speech comprehension research. For instance, on the basis of a simulation run on a purely episodic model, Cutler, Eisner, McQueen, and Norris (2010) argued that a degree of abstraction is necessary to account for how listeners generalize knowledge acquired about a speaker's deviant pronunciations to new words that were not present in the set of words on which the listeners were trained.

Hybrid models have also been used to explain that although many priming experiments show that detailed acoustic characteristics of words play a role in word comprehension, exemplar effects do not replicate across all experiments (e.g., Cooper & Bradlow, 2017; Hanique, Aalders, & Ernestus, 2013; McLennan et al., 2003). In experiments where exemplar effects did not arise, listeners likely relied mostly on abstract representations. Speech comprehension therefore appears to involve both abstract representations and exemplars.

On the basis of the findings in the literature so far, it is far from straightforward to predict under which circumstances exemplar effects arise, largely because of substantial differences in experimental set-ups between studies (see also Goh, 2005). For instance, experiments vary considerably with respect to listeners' tasks, stimulus words, signal-to-noise ratio, the variation forming the basis for the match/mismatch conditions, the number of intervening items between primes and targets (i.e., the lag), the number of items and participants tested, and the dependent measure(s) used. In Table 1.1, we summarize a set of representative studies to illustrate the differences in experimental set-ups as well as the varied circumstances under which exemplar effects arose.

Nevertheless, researchers have put forward generalizations on under which circumstances exemplar effects arise in speech comprehension. The most influential one is the 'time-course hypothesis', formulated by McLennan and Luce (2005). According to the time-course hypothesis, exemplar effects are more likely to arise in late than early processing stages. However, this only holds for exemplar effects of indexical variation (as in McLennan & Luce, 2005), since the opposite pattern (exemplar effects in early but not late processing stages) is predicted for allophonic variation (as in McLennan et al.'s earlier study, McLennan et al., 2003). Existing research should be complemented with replications and new studies to establish under which circumstances exemplar effects exactly arise, and attention should be paid to keeping different experiments comparable among each other.

While the mental lexicon is often assumed as the locus for exemplars (where, in hybrid approaches, exemplars co-exist with abstract representations, such as in McLennan et al., 2003), some researchers have proposed that exemplars are not located in the mental lexicon, but in episodic memory (Cutler et al., 2010; Ramus et al., 2010). Episodic memory refers to the memory of autobiographic events, including all contextual details associated with these events (and is thus not specific to language). The brain regions known to support episodic memory (e.g., the hippocampal complex) are distinct from the brain regions typically associated with language processing (e.g., the bilateral superior temporal lobes).

The locus of exemplars is an important theoretical issue: it is informative about the nature of the mental lexicon. Moreover, if exemplars reside outside of the mental lexicon, this implies that the process of speech comprehension must involve multiple memory systems in parallel. The hypothesis that exemplars are based in episodic memory may also account for the inconsistent findings in the literature in a straightforward manner. If exemplars are based in episodic memory, the occurrence of exemplar effects is expected to specifically relate to how episodic memory functions, and to the extent to which listeners use their episodic memories, which may vary as a function of their task. Although the hypothesis that exemplars are represented in episodic memory rather than in the mental lexicon is discussed in the literature, it has not been tested directly so far.

Table 1.1: Examples of studies testing for exemplar effects in the literature illustrating the varying experimental set-ups and the inconsistency of findings. ‘Task’ refers to the task participants carried out on the targets. LD stands for ‘lexical decision’, ‘old-new’ for old-new judgment, ‘cont. old-new’ for continuous old-new judgment, meaning that primes and targets were intermixed. ‘Dep. variable’ stands for dependent variable, ‘lag’ refers to the number of intervening items between prime and target.

Reference	**Task on targets**	**Dep. variable**	**Variation type**	**Exemplar effects for**	**No exemplar effects for**
Palmeri, Goldinger, and Pisoni (1993)	cont. old-new	accuracy	voice	lags up to 32	lag of 64
Luce and Lyons (1998)	LD; old-new	RT; accuracy	voice	old-new task	LD task
Bradlow, Nygaard, and Pisoni (1999)	cont. old-new	accuracy	voice; rate; amplitude	voice; rate	amplitude
Nygaard, Burt, and Queen (2000)	cont. old-new	accuracy	voice	a-typical productions	typical productions
McLennan, Luce, and Charles-Luce (2003)	LD	RT; accuracy	realization of /t,d/	easy LD	difficult LD
McLennan and Luce (2005)	LD	RT; accuracy	voice; speech rate	difficult LD	easy LD
Mattys and Liss (2008)	old-new	RT; accuracy	voice	dysarthric speech	healthy speech
McLennan and González (2012)	LD	RT; accuracy	voice	foreign-accented speech	native-accented speech
Grohe and Braun (2013)	LD	RT; accuracy	voice; intonation contour	intonation contour	voice
Hanique, Aalders, and Ernestus (2013)	LD	RT; accuracy	voice; vowel reduction	single variation type	multiple variation types
Krestar and McLennan (2013)	LD	RT; accuracy	emotional tone of voice	difficult LD	easy LD
Dufour and Nguyen (2014)	LD	RT; accuracy	voice	low-frequency words	high-frequency words
Theodore, Blumstein, and Luthra (2015)	old-new	RT; accuracy	voice	voice categorization task on primes	passive task on primes
Cooper and Bradlow (2017)	old-new	RT; accuracy	voice; noise	monosyllabic words	bisyllabic words
Strori, Zaar, Cooke, and Mattys (2018)	old-new	RT; accuracy	voice; noise	voice; perceptually integrated noise	perceptually distinct noise

The memory system in which exemplars are represented also affects the kind of information that may be stored in exemplars. If exemplars are represented in episodic memory, listeners are expected to encode many or all types of extra-linguistic information from the speech signal in exemplars, as memory traces in episodic memory are not subject to perceptual filters on the input. In contrast, if exemplars are based in the mental lexicon, restrictions on the types of extra-linguistic information that may be stored in the form of exemplars could be expected. For instance, in this case, the formation of exemplars may adhere to the 'phonetic relevance hypothesis' of Sommers and Barcroft (2006), which states that only acoustic-phonetic variability that affects phonetically relevant properties of the speech signal can impair spoken word identification.

Previous studies suggest that, together with word tokens, listeners may also encode non-linguistic information such as background noise and environmental sounds into long-term memory (e.g., Pufahl & Samuel, 2014; Cooper, Brouwer, & Bradlow, 2015). These findings point to episodic memory as the locus for exemplars. However, there is also evidence for restrictions on the types of variation that may be encoded in exemplars. Bradlow and Pisoni (1999) observed that while exemplar effects arose for variation in speaker voice and in speech rate, the effects were absent for variation in amplitude. This single finding is consistent with the phonetic relevance hypothesis, and therefore compatible with the mental lexicon as the locus for exemplars.

Under the assumption that exemplars are part of the mental lexicon, it is also likely that exemplars only code information that is relevant in a listeners' native language (L1) phonology. The formation of exemplars would then be subject to L1-determined phonological filtering of the input. Morano, Ten Bosch, and Ernestus (in press) addressed this question, and tested non-native listeners on a variation type not relevant to their L1 (Dutch non-native listeners of French on vowel voicing variation type, a variation type common in French but not Dutch). Exemplar effects arose in one of three versions of their experiment, suggesting that listeners' L1 phonology need not restrict the variation types for which exemplar effects arise. The findings of this study with L2 listeners are therefore compatible with episodic memory as the locus for exemplars. Together, previous research on native and non-native listening thus produced somewhat mixed evidence about the variation types that may be encoded in exemplars. More research is necessary to establish if the variation types which may be stored in exemplars point to episodic memory or the mental lexicon as the locus for exemplars.

Research questions

The main research question of this dissertation concerns the nature of exemplars and their role in speech comprehension. Specifically, it tests the hypothesis that exemplars are based in episodic memory rather than in the mental lexicon. This hypothesis will be addressed in two subquestions. The first subquestion concerns the circumstances under which exemplar effects arise in speech comprehension. If exemplars are based in episodic memory, exemplar effects are expected to arise especially when participants rely on their episodic memories. In addition, the occurrence of exemplar effects is expected to relate to properties of episodic memory. If, instead, exemplars are based in the mental lexicon, the occurrence of exemplar effects should not be related to whether participants rely on their episodic memories, or, to properties of episodic memory.

The second subquestion addresses the kinds of extra-linguistic information stored in exemplars (i.e., the *content* of exemplars). If virtually all types of extra-linguistic information from the speech signal can lead to exemplar effects in speech comprehension, this would point to episodic memory as the locus for exemplars. Instead, if the occurrence of exemplar effects is restricted to variation types that are linguistically relevant, or relevant to listeners' native language (L1) phonology, this would point to the mental lexicon as the locus for exemplars.

Outline

This dissertation presents four studies investigating the hypothesis that exemplars are based in episodic memory rather than in the mental lexicon. They all focus on the circumstances under which exemplars play a role in speech comprehension, while the last one also focuses on which types of extra-linguistic information are stored in exemplars.

Although the hypothesis that exemplars are based in episodic memory has not specifically been tested in experiments so far, it is possible to examine the extent to which findings in the literature are in accordance with this hypothesis. Chapter 2 reports a literature review dedicated to this question. It presents four predictions on when exemplar effects are largest under the assumption that exemplars are represented in episodic memory, and tests these predictions against the literature.

Chapter 2 also reports two long-term auditory priming experiments which investigate whether the occurrence of exemplar effects can be linked to properties of episodic memory. Memory traces in episodic memory are not stable; over time, they are integrated into semantic memory. This consolidation process is sped up if during perception, words are processed for meaning rather than for their perceptual properties (probably because links to existing knowledge are emphasized in that case, Craik & Lockhart, 1972). Exemplar effects are only expected if the primes' exemplars are still available to participants by the time they hear the target. Chapter 2 manipulates the tasks participants performed on the primes; it uses perceptual and meaning-based classification tasks. If exemplars are based in episodic memory, exemplar effects should be largest after the perceptual tasks, because these tasks do not enhance primes' exemplars integration into the mental lexicon. In contrast, if exemplars are based in the mental lexicon, no clear effect of prime task is anticipated.

The two experiments in Chapter 2 also addressed whether exemplar effects arise more reliably for tasks that instruct participants to use their episodic memories while processing the targets than tasks that do not instruct participants to do so. If exemplars are represented in episodic memory, exemplar effects should especially arise in a task that instructs participants to use their episodic memory, while the effects should be smaller or absent in a task that does not instruct participants to do so. Chapter 2 thus not only manipulated the tasks on the primes, but also the tasks on the targets. An old-new judgment task (a task in which participants judge whether or not words occurred previously in the experiment) is contrasted with a semantic classification task on the targets. Only the old-new judgment task instructs participants to use their episodic memories.

Like Chapter 2, Chapter 3 compared the occurrence of exemplar effects between an old-new judgment task and a semantic classification task. Chapter 3 did so using an additional experimental measure: it tested for exemplar effects not only in participants' behavioral responses (as in Chapter 2), but also in their EEG brain signals. Because EEG taps more closely into online cognitive processing, EEG may be a more sensitive experimental measure to capture exemplar effects than behavioral measures are.

Chapters 4 and 5 further investigate the circumstances under which exemplar effects arise. Both of these chapters manipulate listening circumstances. If listening circumstances in priming experiments are challenging, listeners may exploit the repetition of word to a large extent, as use of the memory trace of the prime may offer an advantage for processing the target. As such, challenging listening circumstances invite participants' reliance on episodic memory in priming experiments. If exemplars are represented in episodic memory, exemplar effects are therefore expected to arise most clearly under challenging listening circumstances. Under easier listening circumstances, in contrast, exemplar effects are expected to be smaller or absent. In

Chapter 4, challenging listening circumstances were created by adding background noise to stimuli in a lexical decision experiment. This experiment is contrasted with an experiment in which the stimuli were presented without background noise.

Chapter 5 implements challenging listening circumstances by testing different listener populations. It compares the occurrence of exemplar effects in lexical decision tasks between groups of native and non-native listeners. Non-native listening is associated with higher processing costs than native listening (e.g., Borghini & Hazan, 2018). High processing costs invite listeners to use their episodic memories in priming experiments, as reliance on episodic memory offers a processing advantage under challenging listening circumstances. Under the assumption that exemplars are represented in episodic memory, exemplar effects are therefore especially expected to arise for the non-native listeners as a result of their higher processing costs in the lexical decision task.

Chapter 5 also addresses the question which types of surface variation are represented in exemplars. If exemplars are based in episodic memory, no significant differences are expected between different variation types, as all variation can be stored in episodic memory. The mental lexicon, in contrast, likely has constraints on the variation types that may be stored. One variation type we used, variation from speaker voice, is not language-specific. All listeners are used to dealing with this variation type. The other variation type we used results from initial vowel reduction in English. This variation type, in contrast to variation in speaker voice, is language-specific. One group of non-native listeners had Dutch as L1, and was familiar with this type of reduction from their native language phonology, while the other group of non-native listeners had Spanish as L1, and was not. The experiments therefore test if there is a difference in exemplar effects between language-specific and language-unspecific variation types. In addition, the experiments investigate whether in case of a language-specific variation type, there is an effect of being familiar with the variation type from the native language phonology.

Chapter 6 summarizes and discusses the results of these four studies, and addresses how they improve our understanding of what exemplars are. Moreover, it provides recommendations for future research.

The nature and relevance of exemplars in spoken word recognition

Chapter 2

This chapter is a reformatted version of
Annika Nijveld, Martijn Bentum, Louis ten Bosch, and Mirjam Ernestus (submitted). The nature and relevance of exemplars in spoken word recognition.

Abstract

This study investigates whether exemplars reside in episodic memory rather than the mental lexicon. First, we tested this hypothesis against the literature. We formulated four predictions on when exemplar effects should be largest in auditory identity priming experiments. The literature supported some predictions while being inconclusive with others. Second, we tested two of the four predictions directly in two long-term priming experiments. Primes and targets were spoken by the same or a different speaker. In Experiment 1, participants performed old-new judgments on targets (a task requiring episodic memory use), while in Experiment 2, they performed semantic classifications (which does not require episodic memory). In both experiments, participants engaged in familiarization tasks which required perceptual or meaning-related processing of primes. Clear exemplar effects only arose in Experiment 1, and the size of exemplar effects did not depend on the familiarization task. In contrast, in Experiment 2, participants appeared to rely on abstract priming. Post-hoc analyses showed that the amount of priming depended on speaker-specific characteristics. Together, previous and current findings support the hypothesis that exemplars are represented in episodic memory rather than the mental lexicon. This hints at a smaller role for exemplars in speech recognition than has previously been assumed.

Introduction

Hybrid models of spoken word recognition assume that there are two types of memory representations for each word pronunciation: an abstract representation and a cloud of exemplars (e.g., Goldinger, 2007; McLennan et al., 2003; Pierrehumbert, 2002). An abstract representation consists of a sequence of categorical units such as phonemes, and does not contain more detailed information about a word's pronunciation. A cloud of exemplars, in contrast, represents the set of experienced instantiations of a word, and contains token-specific information such as characteristics about the speaker's voice, his speaking rate, and his emotional state.

Most empirical support for the representation of words as clouds of exemplars comes from auditory repetition priming experiments. In these experiments, words are repeated, and participants typically respond more quickly and/or more accurately to the second occurrence (the 'target' token) of a word than to the first (the 'prime' token). Despite some mixed findings in the literature (e.g., Hanique et al., 2013, for an example of null results and discussion of several other null results), it has generally been established that this repetition priming effect is larger when prime and target are acoustically similar, for instance because they are spoken by the same speaker (the match condition). When prime and target are less similar (the mismatch condition), priming is diminished (e.g., Bradlow et al., 1999; Cooper & Bradlow, 2017; Craik & Kirsner, 1974; McLennan et al., 2003; Strori et al., 2018). These 'exemplar effects' suggest that acoustic details of the prime are retained in memory in the form of exemplars, and can therefore influence the recognition of the target.

A question that arises, also in the light of the mixed results in the literature, is where exemplars reside in the mind of the listener: in the mental lexicon, or elsewhere? The locus of exemplars is an important theoretical issue: it is informative about the nature of the mental lexicon, and sheds light on how exemplars affect speech processing, including their role for everyday speech comprehension. Moreover, if exemplars reside outside of the mental lexicon, this implies that the process of speech recognition must involve multiple memory systems in parallel.

Both hypotheses (i.e., exemplars are stored in the mental lexicon or outside of it) are discussed in the literature. McLennan et al. (2003) suggested that clouds of exemplars are represented in the mental lexicon (alongside with abstract representations). According to their model, abstract representations, capturing high-frequent information, dominate speech perception when processing is fast. Exemplars, which capture low frequent information, may come into play when processing is slow - the 'time-course hypothesis' by McLennan and Luce (2005), implemented within the adaptive resonance framework of Grossberg and colleagues (e.g., Grossberg, 1986; Grossberg & Myers, 2000). Relatedly, recent work has proposed an expansion of the mental lexicon as to not only include indexical information in words' representations, but also background noise or environmental sounds (e.g., Cooper et al., 2015; Pufahl &

Samuel, 2014).

The mental lexicon as the locus for exemplars, such as assumed by McLennan et al. (2003), is in line with purely exemplar-based models (which do not assume abstract representations), such as discussed by Goldinger (1998), Johnson (2004), Bybee (2001).

Other authors, instead, have proposed that clouds of exemplars are based in episodic memory, while abstract representations are located in the mental lexicon (e.g., Cutler et al., 2010; Ramus et al., 2010). Importantly, this proposal has the potential of accounting for the mixed results in the literature. This proposal is reminiscent of Goldinger (2007)'s dual, interdependent approach with exemplar-based storage in the hippocampal complex (i.e., the core brain system underpinning episodic memory), and abstract representations in the cortical complex.

Episodic memory refers to the memory of autobiographic events, including all contextual details associated with these events (and is thus not specific to language). If exemplars are located in episodic memory, a number of predictions can be formulated based on logic and on what is known about episodic memory. These predictions concern the circumstances under which exemplar effects in speech comprehension experiments should be largest. We formulated four distinct predictions, which collectively offer a holistic view of the matter that is currently lacking in the literature.

Given that in a number of studies, exemplar effects arose in some experiments, while they did not arise in other experiments, it is likely that in some cases, speech comprehension (mostly) involves abstract representations. If exemplars are based in episodic memory, we expect that exemplars are more likely to play a role in speech comprehension under circumstances where listeners use their episodic memories to a large extent. Prediction I states that exemplar effects are largest in experiments where the task requires participants to use their episodic memories (e.g., when participants are asked to judge if stimuli occurred previously in the experiment, an 'old-new judgment' task). When this is not the case (e.g., in lexical decision tasks), exemplar effects are expected to be attenuated or absent.

Participants may not only rely on their episodic memories because of the task in an experiment, but also because of the listening circumstances. Challenging listening circumstances may encourage participants in repetition priming experiments to benefit from the repetition of words, and to make use of the memory traces of the primes for processing the targets, as doing so may offer a processing advantage. As such, they may be invited to rely on their episodic memories under these circumstances, which promotes the occurrence of exemplar effects (this prediction was also put forward by Mattys & Liss, 2008). Prediction II therefore holds that exemplar effects are larger for less intelligible targets than for clean targets. Chapter 4 of this dissertation also investigates the role of stimulus intelligibility on the occurrence of exemplar effects.

Memory traces in episodic memory show decay, and do therefore not remain available indefinitely. If exemplars are based in episodic memory, exemplar effects are expected to be largest if primes are still available in episodic memory when listeners process targets. According to ‘systems consolidation’ theory, new memories which initially have representations in episodic memory and in some cases also in semantic memory, gradually become stable, long-term memories, which no longer depend on episodic memory (Dudai, 2012; Squire & Alvarez, 1995);(see McClelland, McNaughton, & O’Reilly, 1995; McClelland, 2013, for a computational implementation of this theory termed complementary learning systems). How quickly memories are consolidated depends, among other things, on how well the information can be linked to existing knowledge. If information is perceived as familiar or typical, it can be linked more easily, and thus integrates more quickly (e.g., Kesteren, Rijpkema, Ruiter, Morris, & Fernandez, 2014). Conversely, if information is less typical, it is more likely to *not* integrate quickly, and thus remain available in episodic memory. If exemplars are part of episodic memory, we expect exemplar effects to be largest when, at the presentation of the targets, primes are not yet integrated into (lexical) memory, for instance because they represent uncommon word types (i.e., with low frequencies of occurrence). Prediction III therefore states that exemplar effects are larger for uncommon compared to common word types.

How quickly memory traces can be integrated into semantic memory also depends on how primes are being processed during the encoding phase of an experiment. Meaning-related, but not perceptual processing, is thought to enhance memory consolidation because meaning-related processing emphasizes links between new and existing knowledge (Craik & Tulving, 1975). Material is therefore more likely to still be available in episodic memory if links between the new material and existing knowledge are *not* emphasized. Therefore, if exemplars are part of episodic memory, primes are most likely to still be available to participants if they were processed for perceptual rather than semantic aspects, because perceptual processing does not boost consolidation like meaning-related processing does. Prediction IV states that exemplar effects are largest when participants process primes for perceptual rather than semantic aspects.

We investigated the hypothesis that exemplars are part of episodic memory in two ways. We first verified to what extent the results reported in previous literature are in line with the hypothesis by reviewing existing evidence in favor of or against the four predictions we formulated. Second, we conducted two experiments in which we tested two of our predictions ourselves: that exemplar effects are largest when participants use their episodic memories (Prediction I) and when participants focus on perceptual properties of primes (Prediction IV).

Each experiment consisted of two parts, which both contained a familiarization and a test phase (with primes and targets, respectively). The division into two parts served to keep the lag between primes and targets small, as previous experiments have shown that the likelihood of observing exemplar effects decreases with longer lags (e.g., Hanique et al., 2013). Primes and targets were spoken by a male or female speaker, and prime-target pairs were spoken in the same or the other voice.

We divided the participants in both experiments into three groups that engaged in one of three tasks during the familiarization phases (see Table 2.1 for a summary of the experiments' design), which differ in the type of processing of primes that they require. In the first familiarization task, participants indicated whether words were spoken by the male or the female speaker. This task is perceptual in nature and directs participants' attention to speaker voice, which is the dimension on which primes and targets could differ, and forms the basis of the prime-target match/mismatch condition. A task with a focus on speaker voice could therefore evoke larger exemplar effects than a task without such a focus. In the second familiarization task, participants judged each word's loudness relative to that of the previous word they heard. This task is perceptual as well, but it does not direct participants' attention to speaker voice information. While this task necessitates the creation of a memory trace (essentially, it is a one-back task), it operates on a much shorter time-span (i.e., less than a few seconds, and thus involving short-term memory) than the identity priming in our prime-target pairs (i.e., several minutes); this factor is therefore not expected to significantly boost exemplar effects. In the third familiarization task, participants classified words according to their meaning. This task requires meaning-related processing. Together, our familiarization tasks will show if exemplar effects are larger for perceptual than meaning-related tasks.

In Experiment 1, participants performed old-new judgments during the test phase, which requires participants' use of episodic memory. Experiment 2 was identical to Experiment 1, except that the task in the test phase was semantic classification, for which use of episodic memory is not necessary. Participants classified words according to whether they referred to animate or inanimate objects. We chose this task (rather than, for instance, lexical decision or identification of words embedded in noise) because it allowed us to use the exact same auditory stimuli as in Experiment 1.

Table 2.1: Design of Experiments 1 and 2.

Experiment	**Task familiarization phase**	**Task test phase**
Experiment 1	Voice judgment	Old-new judgment
	Loudness judgment	Old-new judgment
	Semantic classification	Old-new judgment
Experiment 2	Voice judgment	Semantic classification
	Loudness judgment	Semantic classification
	Semantic classification	Semantic classification

Our experiments extend previous work in a number of ways. First, in contrast to a number of existing studies (e.g., Schacter & Church, 1992; Theodore, Blumstein, & Luthra, 2015), we changed tasks between experiments but kept everything else identical, offering a more systematic comparison of tasks.

Second, we systematically varied tasks in both familiarization and test phases. This allows us to not only investigate the influence of familiarization and test tasks separately, but also to examine possible interactions between familiarization and test tasks.

Third, our use of different tokens for primes and targets with natural item-to-item pronunciation variation in both prime-target match and mismatch conditions (following recent work by Hanique et al., 2013, and reported in Chapter4 of this dissertation) improves the ecological validity of our study. In real life speech comprehension, listeners are never faced with two acoustically identical tokens of a word since these do not occur, not even when the two tokens are uttered by the same speaker in a sequence. As such, our study fits into a wider trend in psycholinguistic research to study more realistic speech materials (e.g., Tucker & Ernestus, 2016).

Fourth, our familiarization phase tasks will show what the effect is of participants focusing on the acoustic information that forms the basis of the prime-target match/ mismatch condition by introducing two distinct perceptual tasks. One (the voice classification task) draws participants' attention to the match/mismatch condition (i.e., speaker voice) and one (the loudness judgment task) does not. At the same time, our tasks were designed to avoid participants' use of metalinguistic knowledge, which offers a more precise view on the effect of whether or not tasks are perceptual in nature.

Review of findings reported in the literature

We formulated four predictions on when exemplar effects are largest if exemplars are part of episodic memory. We will now review to what extent these predictions are supported by existing literature. We will discuss findings relevant for our predictions, also when the studies were designed to test different hypotheses.

According to our Prediction I, exemplar effects arise mostly clearly in experiments where the task requires participants to use their episodic memories during test. One previous study supports this prediction. This study aimed to inspect if exemplars are not only part of some form of memory, but also affect on-line processing. Exemplar effects arose when participants performed old-new judgments during test, while exemplar effects did not arise when participants engaged in lexical decisions on the same stimuli (Luce & Lyons, 1998). The authors concluded that on-line processing (as is necessary for the lexical decision task) likely relies mostly on abstract representations. Goldinger (1996) tested for exemplar effects in an old-new judgment task and a task in which participants needed to select the best semantic associate for the same words. He established exemplar effects for both tasks; this study does thus not support Prediction I.

Prediction II states that exemplar effects are larger for less intelligible stimuli. Most available evidence is in line with this prediction: in studies which manipulated listening difficulty while keeping the task constant, larger exemplar effects arose for the less intelligible materials (e.g., dysarthric speech, foreign-accented speech and noisified speech in Mattys & Liss, 2008; McLennan & González, 2012; Saldaña, Nygaard, & Pisoni, 1996, respectively). A null result was found by Theodore et al. (2015): exemplar effects did not arise for targets presented in clean, nor for targets embedded in background noise. The studies of Mattys and Liss (2008) and McLennan and González (2012) were designed to manipulate processing speed to test the time-course hypothesis of McLennan and Luce (2005), while the main goal of the study by Saldaña et al. (1996) was to investigate whether detailed cross-modal linguistic information is stored into memory. The main goal of the study by Theodore et al. (2015) was to investigate attention allocation in various familiarization tasks during encoding; this study will again be discussed below.

According to Prediction III, exemplar effects are larger for stimuli which are not consistent with prior knowledge. Studies by Goldinger (1996, 1998), Dufour and Nguyen (2014), and Dufour, Bolger, Massol, Holcomb, and Grainger (2017) found support for this: exemplar effects were larger for low frequency words. The study by Dufour and Nguyen (2014) was designed to investigate the time-course hypothesis (McLennan

& Luce, 2005), while the study by Dufour et al. (2017) aimed to investigate which stages of spoken word recognition, as captured by EEG, are mostly affected by listeners' use of exemplars. Goldinger (1996) investigated a number of issues, such as tasks (as discussed in the context of Prediction I above and of Prediction IV below) and the delay between prime and target; the frequency effect was not one of the main manipulations, and was reported as a post-hoc analysis in Goldinger (1998). The goal of Goldinger (1998) was to test an episodic model against data from a shadowing experiment. Another study investigated the influence of perceived typicality of surface characteristics on exemplar effects, and varied the way in which stimuli were produced. Larger exemplar effects for stimuli produced in less typical ways (i.e., at a slow speech rate, high amplitude, or loud vocal effort) were found (Nygaard, Burt, & Queen, 2000). Prediction III is thus also validated by the literature on exemplar effects.

Finally, Prediction IV states that exemplar effects are larger when participants engage in perceptual rather than meaning-related processing of primes (imposed by the task in the familiarization phase). A number of studies have manipulated tasks in the familiarization phase (e.g., Goldinger, 1996; Naveh-Benjamin & Craik, 1995; Schacter & Church, 1992; Scheffert, 1998; Theodore et al., 2015). Naveh-Benjamin and Craik (1995) investigated exemplar effects for spoken and written words for participants of different age groups, Schacter and Church (1992) studied the effect of various familiarization tasks on both implicit priming and recognition memory, and Scheffert (1998) and (Theodore et al., 2015) investigated the occurrence of exemplar effects after various familiarization tasks. In perceptual tasks, participants made judgments about words' speaker voices (e.g., Goldinger, 1996; Naveh-Benjamin & Craik, 1995; Theodore et al., 2015), utterance clarity (e.g., Scheffert, 1998), or pitch (e.g., Naveh-Benjamin & Craik, 1995; Schacter & Church, 1992). An 'intermediate' task required participants to make judgments about words' initial phonemes (Goldinger, 1996). In meaning-related tasks, participants made judgments about words' membership to semantic (e.g., Schacter & Church, 1992) or syntactic categories (Goldinger, 1996; Theodore et al., 2015), or their numbers of meanings (e.g., Scheffert, 1998).

Two studies found that exemplar effects varied as a function of familiarization task, such that more perceptual tasks generally evoked larger exemplar effects (Goldinger, 1996; Theodore et al., 2015). The task showing the largest exemplar effects was most perceptual in nature (speaker voice classification), but also required participants to focus on the variation that formed the basis for the prime-target match/mismatch condition: speaker voice. It is unclear how this influenced the exemplar effects in these studies. Furthermore, both of these studies used metalinguistic tasks: the less perceptual initial phoneme classification task in Goldinger (1996) and the non-perceptual syntactic category classification task in Goldinger (1996) and Theodore et al. (2015). The activation of participants' metalinguistic knowledge may evoke additional processes, and therefore interfere with the occurrence of exemplar effects. For instance, metaphonological tasks have been shown to not only activate brain areas specialized in speech processing, but also recruit areas involved in visual processing (e.g., Booth et al., 2002). Hence, it is not entirely clear which factor(s) exactly increased or decreased exemplar effects in these studies. This leaves it undecided whether Prediction IV is supported by existing literature.

Findings from a number of studies are of relevance for more than one prediction; the findings of most of these were discussed in the context of several predictions above. Exceptions are Schacter and Church (1992) and Church and Schacter (1994), which we have not yet discussed. In these studies, exemplar effects arose when participants engaged in the completion of word stems or the identification of acoustically modified words. These tasks do not require the use of episodic memory (of relevance for Prediction I), and should therefore produce minimal exemplar effects. However, the experiments included degraded stimulus materials (of relevance for Prediction II), which should increase the exemplar effects. It seems that exemplar effects arose as a consequence of these degraded stimuli.

Taken together, existing literature never contradicts our predictions. Rather, it either supports them (i.e., Predictions II and III, which were each validated by several studies, and Prediction I, which received support from a single study), or the literature is indecisive about our predictions (i.e., Prediction IV, for which relevant studies may have been confounded). Hence, it is quite possible that exemplars are represented in episodic memory. To further investigate this hypothesis, we tested Predictions I and IV in more detail in two experiments.

Experiment 1

Method

Participants

Ninety-three highly educated native speakers of Dutch, aged between 18 and 26 years (mean: 21 years), received a small monetary compensation to participate in the experiment. Eighteen were male, 15 were left-handed. We divided the participants equally over three familiarization task groups (31 participants in each group). Participants who took part in this experiment did not participate in Experiment 2.

Materials

The stimuli consisted of 72 bisyllabic Dutch nouns (e.g., lepel ‘spoon’ and schaduw ‘shadow’), of which 36 served as experimental words and 36 as fillers (see the Appendix). Experimental words’ log-transformed frequencies of occurrence ranged between 0.70 and 4.02 (2.33 on average), while the fillers’ log-transformed frequencies ranged between 0.00 and 4.40 (2.25 on average; counts from the CELEX Dutch lexical database, Baayen, Piepenbrock, & Gulikers, 1995).

We recorded eight tokens of each word with a male and a female speaker of Dutch in a sound-attenuating booth at a 44.1 kHz sampling rate. The speakers read the words from a list and were instructed to keep their intonation as constant as possible. For the fillers, which occurred once in the experiment, we selected the first good-sounding token from each speaker, while for the experimental words, which occurred twice, we selected the first two good-sounding tokens from each speaker.

We equalized our selected tokens at an average loudness of 70 dB (in Praat Boersma & Weenink, 2018). This energy level was measured on the entire stimulus. Despite this dB equalization, the resulting stimuli are likely to differ in perceived loudness, for at least two reasons. First, two different speech sounds played at the same dB level may differ in the perceived loudness due to the spectral differences between them, as the dB scale does not take into account the (non-linear) frequency-dependent sensitivity of the human hearing system (i.e., it is not corrected for properties of the auditory system). Second, the 70 dB calibration holds as an average across the entire signal, and so two tokens may still differ in the way they distribute the energy along the signal.

The selected prime and target tokens of the experimental words spoken by the male speaker had an average duration of 596 ms (standard deviation (*SD*) = 106 ms), while those spoken by the female speaker were 630 ms on average (*SD* = 105 ms), which is a statistically significant difference (t (71) = -2.12, $p < .05$). Participants in the experiment, however, did not hear all four tokens but only two of each experimental word. On a randomly chosen subset of tokens that a participant could hear, word duration did not differ between our two speakers (male speaker: mean = 611 ms, *SD* = 95 ms; female speaker: mean = 616 ms, *SD* = 98 ms; t (35) = -0.25, p = .8)

Even though our speakers were both normal, healthy talkers of a standard language variety of Dutch, we carried out a rating experiment to explore possible differences in intelligibility between our two speakers. Ten subjects (who did not participate in the main experiments; three males) rated the clarity of the enunciation of each experimental prime token. Each subject heard every experimental words once. Half of these were spoken by the male speaker and half by the female speaker; we counterbalanced across participants which speaker produced which word. Participants made their judgments on a six-point scale (where a score of one meant that the pronunciation of the word was very unclear, while a score of six meant that it was very clear). The average ratings (male speaker: 5.0, *SD* = 1.09; female speaker: 4.25, *SD* = 1.36) showed that both speakers were easily understandable, but the male speaker was rated as significantly more intelligible than the female speaker (*t* (179) = 7.29, *p* < .005). In addition, we assessed differences in intelligibility by comparing how often each speaker received the highest score of six. This was more often the case for the male speaker (78 times) than for the female speaker (34 times). A generalized linear mixed effects model showed that this difference is statistically significant as well ($\hat{\beta}_{\text{female speaker}}$ = -2.0, *z* = -5.5, *p* < .001). We therefore conclude that even though both speakers' utterances are sufficiently intelligible, the male speaker's pronunciations are uttered relatively more clearly.

We also assessed the similarity of prime and target tokens produced by the same speaker in two ways, again comparing the two speakers. First, we compared absolute duration differences for the two speakers. We found that for all 36 prime-target pairs, the male speaker's pair members differed on average 37 ms from each other (*SD* = 30 ms), while those of the female speaker differed on average 44 ms from each other (*SD* = 37 ms). This difference is not statistically significant (*t* (35) = -1.03, *p* = .30). For a randomly chosen subset of these pairs that a given participant could hear, the male speaker's pair members had an average absolute duration difference of 27 ms (*SD* = 24 ms) and the female speaker's pair members 54 ms (*SD* = 45 ms), a difference that is also not statistically significant (*t* (8) = -1.38, *p* = .20).

Second, we had six phonetically trained listeners (two male) judge the similarity of prime-target pair tokens produced by both speakers. They did so on a six-point scale, where a score of one meant that the two tokens differed maximally from each other, while a score of six meant that they sounded identical. The male speaker's prime-target pair tokens were on average rated as more similar (4.37, *SD* = 1.52) than those of the female speaker (3.76, *SD* = 1.53). This difference was statistically significant (*t* (215) = 5.31, *p* < .005). We also determined for each word which speaker obtained the highest rating. This speaker was given a score of 1, while the other speaker obtained a score of 0 (both speakers were attributed 0 in case of the same rating). Out of the 432 ratings in total (36 prime-target pairs x 2 speakers x 6 raters), the male speaker's pairs obtained a score of one 114 times, while the female speaker's pairs obtained a score of one 59 times, which is a statistically significant difference according to an exact sign test (*p* < .001). So, although the two speakers did not differ in the absolute duration difference between their prime-target pair members, the male speaker's primes and targets were perceived as relatively more similar than those produced by the female speaker.

In the main experiment, a prime and target appeared in the same or a different voice, which yields four possible combinations of speaker voice (male - male, female - female or female - male, male - female). We created three pseudo-randomized master lists, in which we divided the prime-target pairs equally over the four voice combinations. In each of the four experimental phases (familiarization - test - familiarization - test) of each list, half of the experimental words were spoken by one speaker and half by the other speaker. Half of the fillers in the experiment - occurring in the test phases only - were spoken by each speaker. We ensured that no more than three experimental words or fillers and no more than three trials with the same speaker voice occurred consecutively. Each test phase started with at least two fillers.

For each master list, we created three additional lists with the same word order, in each of which each prime-target pair represented a new combination of speaker voice. This way, every prime-target pair represented each speaker voice combination exactly once in the set of every master list.

Procedure

We divided the experiment into two parts, where each part consisted of a familiarization phase (with 18 experimental words appearing as primes), and a test phase (with the same 18 words appearing as targets, in addition to 18 fillers). The two familiarization phases and two test phases together thus yielded a total of 108 trials. The division of the experiment in two parts served to keep the lag between primes and targets small (primes and targets were separated by 26 trials on average, while the maximum lag was 49), because long lags (> 100 trials) have been shown to decrease the likelihood of finding exemplar effects (Hanique et al., 2013).

For the familiarization phases, the procedure differed for the three groups of participants. We asked the first participant group to classify each word according to whether it was spoken by a man or a woman ('man'-responses with a press on a button box with the index finger of the dominant hand; 'woman' with the non-dominant hand). We instructed the second group of participants to make judgments about the loudness of the auditory stimulus presented in the current trial relative to the previous trial. Participants could either indicate that the word had the same loudness as the previous word (with their dominant hand) or that they thought it was louder or less loud. On the first trial, we asked participants to not give a response but just listen to the stimulus. We asked the third group of participants to classify words according to whether their referents can typically be found outdoors (with their dominant hand) or indoors. For example, *iglo* ‘igloo’ is something typically encountered outdoors, while *lepel* ‘spoon’ is typically encountered indoors. Most of the 36 experimental words clearly fell into one of these categories (27 words), whereas others (nine words, e.g., *emmer* ‘bucket’ and *schaduw* ‘shadow’) were more neutral. We instructed the participants in all groups to also pay attention to the words themselves because that would be of importance later in the experiment (such that they would also encode information about word identity). All familiarization tasks were unspeeded (i.e., there was no time limit for responding).

For the test phases that followed the familiarization phases, we instructed all participants to judge words as quickly and accurately as possible as ‘old’ if they had occurred before in the experiment (this was the case for the experimental words, half of the trials) and as ‘new’ if they had not (which was the case for the fillers). ‘Old’ judgments were given with the dominant hand.

After completing the first part of the experiment (i.e., after one familiarization and one test phase), participants engaged in a distractor task for four minutes, in which they solved arithmetic problems (such as the subtraction 160 - 113) on a sheet of paper. Its purpose was to clear the words of the first part of the experiment from participants’ working memories (to some extent), and to let some time pass between the two parts. After the distractor task, the second part of the experiment was initiated and participants were informed that they would have to do the same tasks again, and that they could forget about the first part of the experiment.

We tested participants individually in a sound-attenuating booth. We presented stimuli over closed headphones at a comfortable listening level with E-prime 2.0 experimental software (Psychology Software Tools, Pittsburgh, PA). On each trial, participants saw a fixation asterisk for 100 ms on a computer screen, and after a pause of 200 ms (blank screen), one auditory stimulus (i.e., a word) was presented. After this auditory stimulus, we showed the response labels on the left and right on the screen to remind participants which button on their box corresponded to which response (font: Courier New; size: 30; the left or right position of the labels on the screen was adjusted to the participants’ handedness). In the familiarization phases, we showed the words *man* ‘male’ and *vrouw* ‘female’ (first group), *even hard* ‘same loudness’ and *harder/zachter* ‘louder/less loud’ (second group), *binnen* ‘indoors’ and *buiten* ‘outdoors’ (third group), while in the test phases, we showed the words *oud* ‘old’ and *nieuw* ‘new’. The next trial was started after the response; there was no time limit. The session lasted approximately 10 minutes.

Analyses

We analyzed participants' accuracy of experimental words in the test phases (i.e., the targets) by means of generalized mixed effects regression models in `R` statistical software (RCoreTeam, 2018) with the binomial link function of the `lme4` package (Bates, Mächler, Bolker, & Walker, 2015). We tested for the influence of *Speaker match* (which reflected whether a prime and target were spoken by the same speaker or not), *Familiarization task* (voice judgment, loudness judgment, or semantic classification), and their interaction.

In addition, we tested for effects of a number of control predictors: log-transformed Word duration, log-transformed *Word frequency*, *Trial number*, *Part* of the experiment (part 1 or 2), *Speaker* (male or female), and *Lag* in trials between prime and target. We used *Word* and *Participant* as crossed random effects, and tested for random slopes for the predictors of interest and their interactions (by-subject and by-word). Inclusion of these random slopes did not result in significant improvements of the statistical model in any of our analyses. We did not test for random slopes of control predictors for three reasons: first, we had no experimental hypotheses about our control predictors, second, doing so enhances the likelihood of overfitting of our model to this particular dataset (which decreases the generalizability of our findings), and third, doing so increases the chances of model convergence issues. We only included significant effects and interactions in the model (i.e., whose absolute *t*-values exceeded 1.96, which implies $p < .05$ for this type of data, as the many observations our datasets approach a normal distribution), as well as simple effects of predictors appearing in significant interactions.

We analyzed log-transformed RTs (measured from word onset) to targets which received correct 'old' responses with mixed effects regression models. We restricted the analysis to targets whose primes had received correct responses in the voice classification task (the other two tasks - loudness and indoor/outdoor judgment - were more subjective). In addition, we excluded trials whose RTs were more than two standard deviations apart from the grand mean. For the best model, we removed data points with standardized residuals exceeding 2.5 standard deviation units and refitted the model. We used the same predictors as in the analysis of the accuracy scores, except that we also included log-transformed reaction times to the prime (*RT prime*) and to the preceding trial (*RT preceding trial*) as control predictors.

Since we could not reliably compare RTs in the familiarization and test phases due to large differences in task requirements (e.g., unspeeded versus speeded), we did not examine whether targets elicited faster responses than primes to test for repetition priming. Repetition priming could also not reliably be established by comparing RTs to primed targets versus unprimed fillers in the test phases, because targets and fillers were responded to with different hands (the dominant and the non-dominant one, respectively).

Previous studies found no indication that exemplar effects differ between normal talkers (e.g., Hanique et al., 2013). As such, we had no a-piori experimental hypotheses about larger exemplar effects for either of our talkers, and our experiments were not specifically designed to test for this. However, we still wished to explore whether *Speaker Match* interacted with *Speaker* in post-hoc analyses, as preceding literature has frequently established restrictions of exemplar effects to specific cases (see e.g., Nygaard et al., 2000, and the literature review above). It may well be that talkers' indexical idiosyncrasies affect the occurrence of exemplar effects. More specifically, we established that our speakers differ on at least two potentially relevant dimensions: their intelligibility (see also Prediction II) and their prime-target token similarity.

Table 2.2: Participants' behaviour in the familiarization phases of Experiments 1 and 2 split by familiarization task. *SE*s of the means for the RTs are given in brackets. For the voice task, probability correct is given with 95% Wilson score confidence intervals in brackets. The loudness task and semantic classification tasks are more subjective in nature; for these tasks, we report the average proportion of response agreement for participants receiving the same experimental lists.

		RTs (ms)	**Probability correct**
Experiment 1	Voice judgment	987 (14)	0.99 (0.99-1.00)
	Loudness judgment	1368 (12)	0.63 agreement
	Semantic classification	1501 (17)	0.79 agreement
Experiment 2	Voice judgment	913 (7)	0.99 (0.99-1.00)
	Loudness judgment	1687 (21)	0.55 agreement
	Semantic classification	1291 (12)	0.76 agreement

Results and discussion

Participants' performance in the familiarization phases is summarized in Table 2.2. We report RTs for all familiarization tasks. Probability correct is reported in case answers could be scored as correct or incorrect (i.e., in the voice classification familiarization task). We report the extent to which participants agreed in their responses for the two more subjective tasks (loudness judgment and indoor-outdoor judgment). For the indoor-outdoor task, two thirds of the words were not subjective, and therefore, an agreement of about 0.75 is expected for that task. For the loudness judgment task, agreement that is systematically different from chance would be agreement that is higher than the upper limit of the 95% Wilson score confidence interval (Wilson, 1927) for a probability of 0.5 correct, which amounts to 0.531 for our 1116 observations for this task.

For the test phases, participants' performance is summarized in Table 2.3. We found that participants were on average 90% correct in the old-new judgment task (to experimental and filler words; *SD* = 7%). Response accuracy on just the experimental ('old') words was 88% (*SD* = 11%) on average.

Response accuracy could only be predicted by *Word frequency* ($\hat{\beta}$ = 0.42, z = 4.2, p < .001): words with a higher frequency of occurrence received more incorrect 'old' responses. This effect has been reported before (in the literature on visual old-new judgment, e.g., Glanzer & Adams, 1985), and reflects that words with high frequencies of occurrence are inherently more familiar than words with low frequencies of occurrence. This makes them more likely to be called 'old', even when they did not appear in the familiarization phase. Our predictors of interest *Speaker Match*, *Familiarization task*, and their interaction did not reach statistical significance.

Table 2.3: Participants' behaviour to experimental words in the test phases in the old-new judgment task in Experiment 1, split according to *Familiarization task* and *Speaker match*. For the mean RTs to correctly classified targets, *SE*s are given in brackets, and for accuracy, probability correct is given with the lower and upper boundaries of 95% Wilson score confidence intervals in brackets. 'Match' and 'mismatch' refer to the speaker match- and mismatch conditions. 'Voice judgm.', 'Loudness judgm.', and 'Semantic class.' respectively indicate the voice judgment, loudness judgment, and semantic classification familiarization tasks.

	RT (ms)		**Probability correct**	
	Match	Mismatch	Match	Mismatch
Voice judgm.	982 (11)	997 (12)	0.90 (0.87-0.92)	0.88 (0.85-0.91)
Loudness judgm.	965 (10)	997 (11)	0.83 (0.79-0.86)	0.79 (0.75-0.81)
Semantic class.	964 (11)	968 (10)	0.95 (0.93-0.97)	0.95 (0.93-0.97)

The post-hoc analysis with the interaction between *Speaker Match* and *Speaker* shows a statistically significant interaction between *Speaker* and *Speaker Match* ($\hat{\beta}_{\text{female speaker}}$ = -0.35, z = 2.3, $p < .05$; $\hat{\beta}_{\text{speaker match}}$ = -0.43, z = -2.7, $p < .01$; $\hat{\beta}_{\text{female speaker}}$ x $_{\text{speaker match}}$ = 0.45, z = 2.0, $p < .05$). To interpret this interaction, we split the data according to speaker and fitted separate statistical models (with the same predictors, apart from the simple and interaction effects of *Speaker*) to these datasets (see Table 2.5).

Table 2.4: Final statistical model predicting the RTs of correct responses to experimental words in the test phases of Experiment 1. The intercept represents *Experiment part 1*, *Male speaker*, and *Speaker mismatch*.

Fixed effects	$\hat{\beta}$	t
Intercept	3.96	17.7
Experiment part 2	-0.04	-7.6
Word duration	0.31	10.1
RT preceding	0.10	8.2
RT prime	0.04	4.0
Female speaker	-0.02	-3.0
Speaker match	-0.02	-3.3
Random effects		*SD*
Word	intercept	0.04
Participant	intercept	0.10
Residual		0.14

Experimental words in the *Speaker match* condition only received more correct responses (90% correct on average, *SD* = 13%) than experimental words in the *Speaker mismatch* condition (86%, *SD* = 17%) when they were produced by the male speaker (for words produced by the female speaker, *Speaker match* and *Speaker mismatch* both yielded 89% correct, *SD* = 13% and 14%, respectively). The post-hoc analysis thus revealed an exemplar effect in response accuracy, but only for the male speaker.

We analyzed RTs of correct responses to experimental words in the test phases after removing outliers (i.e., data points that exceeded two standard deviations from the grand mean, 4% of the data). These RTs ranged from 604 to 1609 ms and were 978 ms on average (*SD* = 199 ms).

Table 2.4 presents the final statistical model for the RTs. We found effects of the control predictors *Experiment part*, *Word duration*, *RT preceding*, *RT prime*, and *Speaker*, indicating quicker responses to words in the second part of the experiment, words with shorter durations, words whose preceding trials or primes received quicker responses, and words produced by the female speaker.

Table 2.5: Post-hoc statistical models predicting the accuracy of responses to experimental words in the test phases of Experiment 1, split by speaker. The intercept represents *Speaker mismatch*.

	Male speaker			**Female speaker**		
Fixed effects	$\hat{\beta}$	z	p	$\hat{\beta}$	z	p
Intercept	-3.30	-10.3	< .001	-3.30	-8.7	< .001
Word frequency	0.51	4.6	< .001	0.35	2.6	< .01
Speaker match	-0.44	-2.7	< .01	-0.01	-0.1	< 1
Random effects	*SD*			*SD*		
Word (intercept)	0.29			0.53		
Participant (intercept)	0.86			0.90		

More importantly, we observed an effect of the predictor of interest *Speaker match*: responses to target words spoken in the same voice as their primes were quicker (970 ms on average, *SD* = 201) than responses to words spoken in the other voice as their primes (986 ms, *SD* = 197). We did not find an interaction between *Speaker match* and *Familiarization task*. Our post-hoc analysis with *Speaker match* and *Speaker* showed no significant interaction between these two predictors either. Hence, in the RT data, we found exemplar effects that seem independent of speaker and familiarization task.

Given the repetition of tasks in the two parts of the experiment, participants may show different behaviour in the two parts. For example, they may pay closer attention to words in the second familiarization phase because they know old-new judgment will follow. Since this could affect the strength of exemplar effects, we checked for an interaction between *Speaker match* and *Experiment part*. No such interaction arose. Hence, we found no indication that exemplar effects differ between the two parts of the experiment.

In short, Experiment 1 showed exemplar effects, which did not differ between familiarization tasks. However, our post-hoc analysis of accuracy scores showed that exemplar effects may differ between speakers. To investigate the effect of test task in Experiment 2, we replaced the old-new judgment task by a task that does not require participants' use of episodic memory: animacy judgment.

Experiment 2

Method

Participants

Ninety-three highly educated subjects (from the same participant pool as the subjects in Experiment 1), aged between 18 and 28 years (mean: 21 years) participated; 23 were male and ten were left-handed. These participants were divided over three equal groups (31 participants in each group), which engaged in either voice judgment, loudness judgment, or semantic classification during familiarization.

Materials, procedure, and analyses

The materials, procedure, and analyses were identical to those of Experiment 1, with a few exceptions. Most importantly, we changed the task in the test phases: we now instructed participants to judge words' animacy, a semantic task that does not encourage participants to use their episodic memories. All 36 experimental words (thus including the items of both the match- and mismatch conditions) in the experiment referred to inanimate objects. Of the 36 fillers, 12 also had inanimate referents while 24 had animate referents.

Moreover, immediately after each test phase, we presented participants with visual stimuli for which they had to make old-new judgments. The reason for this is that we instructed participants to pay attention to the words themselves (in addition to carrying out the main task) in the familiarization phases. In Experiment 1, this was justified because of the old/new judgment test task. The semantic test task used in Experiment 2 does not similarly reward participants for having paid attention during familiarization. We feared that participants would not pay attention to the familiarization words again in the second part of Experiment 2 once they noticed that there was no need to do so.

The visual stimuli consisted of the 36 experimental words of the test phase in addition to 36 new filler words (provided in the Appendix). These fillers resembled the experimental words: they were also bisyllabic nouns, and their frequencies of occurrence did not show a statistically significant difference with those of the experimental words (*t* (35) = -0.42, *p* = .7). Six of the filler words referred to animate entities. Trials in the visual old-new judgment task consisted of the presentation of a grey screen (600 ms) followed by the presentation of the visual word in the centre of the screen (font: Courier New; fontsize: 30). We showed the response labels *oud* ('old') and *nieuw* ('new') on the screen in the same font during the entire trial. 'Old'-responses were given with the dominant hand. The next trial was initiated after the response; there was no time limit. The complete experimental session lasted about 15 minutes.

Results and discussion

Participants' behaviour in the familiarization phase is presented in Table 2.2 (RTs, and error rates in case of the voice task). In the test phase (participants' behaviour summarized in 2.6), performance was at ceiling: participants were on average 98% correct on experimental and filler words (*SD* = 1%), and 99% correct on experimental words only (*SD* = 2%). This was also reflected by the fact that accuracy could not be predicted by any of our predictors in a statistical analysis. In our post-hoc analysis however, we found an interaction between *Speaker match* and *Speaker* ($\hat{\beta}_{\text{female speaker}}$ = -0.80, *z* = -1.7, *p* = .10; $\hat{\beta}_{\text{speaker match}}$ = -0.80, *z* = -1.7, *p* = .09; $\hat{\beta}_{\text{female speaker}}$ x $_{\text{speaker match}}$ = 1.77, *z* = 2.8, *p* < .01). Although these betas should be interpreted with caution (because of the ceiling performance), they suggest an exemplar effect for the male speaker, and a reversed effect for the female speaker (for the male speaker data: *Speaker match* condition 99% correct (*SD* = 3%), *Speaker mismatch* condition 98% correct (*SD* = 5%); for the female speaker data: *Speaker match* condition 98% correct (*SD* = 6%), *Speaker mismatch* condition 99% correct (*SD* = 3%). We could not further analyze this interaction by splitting the data by speaker, given that the models fitted to these datasets failed to converge.

After removing outlying RTs (i.e., data points exceeding two standard deviations from the grand mean; 5% of the data), RTs of correct responses to experimental words in the test phases ranged from 576 to 1529 ms and were 928 ms on average (*SD* = 184 ms).

Table 2.7 presents the final statistical model fitted to the RTs to correctly classified targets. We found statistically significant effects of *Experiment part*, *Word duration*, *Lag*, *Speaker* and *RT prime* (indicating quicker responses to words in part two of the experiment, to words with shorter durations, words separated further from their primes, words produced by the female speaker, and to words whose primes received

Table 2.6: Participants' behaviour to experimental words in the test phases in the semantic classification task in Experiment 2, split according to *Familiarization task* and *Speaker match*. For the mean RTs to correctly classified targets, *SE*s of the means are given in brackets, and for accuracy, probability correct is given with the lower and upper boundaries of Wilson score 95% confidence intervals in brackets. 'Match' and 'mismatch' refer to the speaker match- and mismatch conditions. 'Voice judgm.', 'Loudness judgm.', and 'Semantic class.' respectively indicate the voice judgment, loudness judgment, and semantic classification familiarization tasks.

	RT (ms)		**Probability correct**	
	Match	Mismatch	Match	Mismatch
Voice judgm.	982 (21)	997 (23)	0.99 (0.98-1.00)	0.99 (0.98-1.00)
Loudness judgm.	965 (20)	997 (23)	0.99 (0.97-0.99)	0.98 (0.97-0.99)
Semantic class.	964 (21)	968 (19)	0.97 (0.95-0.98)	0.99 (0.97-0.99)

quick responses). We will return to the effect of lag in the General Discussion. More importantly, we obtained no simple effect of our predictor of interest *Speaker match*, nor did we find an interaction between *Speaker match* and *Familiarization task*.

Our post-hoc analysis, in which we tested for differences in exemplar effects between our speakers, however, showed an interaction between *Speaker match* and *Speaker* ($\hat{\beta}_{\text{female speaker}}$ = -0.03, t = -3.8; $\hat{\beta}_{\text{speaker match}}$ = -0.01, t = -1.5; $\hat{\beta}_{\text{female speaker x speaker match}}$ = 0.03, t = 2.6). To investigate this interaction in more detail, we performed additional analyses on the data split by speaker. Table 2.8 shows the statistical models for the two speakers separately. These analyses reveal an exemplar effect for the male speaker: For targets spoken by the male speaker, participants responded significantly more quickly to words repeated in the same voice (925 ms on average, *SD* = 183) than in the other voice (940 ms, *SD* = 188). For the female speaker, instead, the effect was reversed: Words repeated in the other voice received significantly quicker responses (915 ms, *SD* = 181) than words repeated in the same voice (932 ms, *SD* = 181). This is the same pattern we observed for participants' accuracy scores.

As in Experiment 1, we checked for an interaction between *Speaker match* and *Experiment part*. There was no such interaction; exemplar effects did thus not seem to differ depending on which part of the experiment participants were in.

Table 2.7: Final statistical model predicting the RTs of correct responses to experimental words in the test phases of Experiment 2. The intercept represents *Experiment part 1* and *Male speaker*.

Fixed effects	$\hat{\beta}$	*t*
Intercept	4.98	24.2
Experiment part 2	-0.03	-5.5
Word duration	0.28	9.1
Lag	-0.003	-10.9
RT prime	0.02	2.3
Female speaker	-0.02	-3.1
Speaker match	*n.s.*	*n.s.*
Random effects		*SD*
Word	intercept	0.04
Participant	intercept	0.11
Residual		0.14

In short, while our main analyses yielded null results for *Speaker Match* in Experiment 2, our post-hoc analyses showed exemplar effects that were restricted to the words produced by the male speaker in both response accuracy and response times (i.e., we obtained interactions between *Speaker Match* and *Speaker*). Reversed exemplar effects arose for the female speaker in Experiment 2 in response accuracy and reaction times. Any exemplar effects for accuracy in this experiment should, however, be interpreted with caution due to ceiling performance.

Our main analyses of the two experiments showed exemplar effects in the RTs of Experiment 1 only (while null results for *Speaker Match* were established in accuracy of Experiment 1 and RTs and accuracy of Experiment 2). To investigate if the two experiments truly differ in the occurrence of exemplar effects, we ran an additional analysis on the RT data of the two experiments combined, in which we tested for an interaction between *Speaker match* and *Experiment* (by adding the interaction term to the richest RT model, i.e., that of Experiment 1). We did not perform such an analysis on the combined accuracy data because of the ceiling performance in Experiment 2. The combined analysis of the RTs showed a statistically significant interaction between *Speaker match* and *Experiment* ($\hat{\beta}_{\text{speaker match}}$ = -0.02, *t* = -3.3; $\hat{\beta}_{\text{experiment 2}}$= -2.05, *t* = -11.5; $\hat{\beta}_{\text{speaker match}}$ x $_{\text{experiment 2}}$ = 0.02, *t* = 2.6), confirming that

Table 2.8: Post-hoc statistical models for RTs of correct responses to experimental words in the test phases of Experiment 2, split by speaker. The intercept represents *Experiment part 1* and *Speaker mismatch*.

	Male speaker		**Female speaker**	
Fixed effects	$\hat{\beta}$	t	$\hat{\beta}$	t
Intercept	4.39	15.4	5.26	17.2
Experiment part 2	-0.03	-3.8	-0.03	-3.5
Word duration	0.37	8.5	0.23	5.0
Lag	-0.003	-7.6	-0.003	-7.9
RT prime	0.03	2.4	0.23	1.9
Speaker match	-0.01	-2.1	0.02	2.5
Random effects	*SD*		*SD*	
Word	(intercept)	0.04		0.04
Participant	(intercept)	0.10		0.11
Residual		0.13		0.14

exemplar effects were larger in the RTs of Experiment 1 than in Experiment 2.

General discussion

This study investigated the nature and role of exemplars in speech comprehension. We tested the hypothesis that exemplars are represented in episodic memory rather than in the mental lexicon. This idea has been proposed before (e.g., Goldinger, 2007; Ramus et al., 2010), but despite its major theoretical implications, has not been tested directly (although relevant circumstantial evidence was reported by e.g. Cooper & Bradlow, 2017; Dufour et al., 2017; Pufahl & Samuel, 2014). The hypothesis that exemplars are represented in episodic memory is potentially powerful in explaining mixed previous results in the literature in a straightforward manner.

The most influential alternative account for these mixed results is the ‘time-course hypothesis’, formulated by McLennan and Luce (2005). According to the time-course hypothesis, exemplar effects are more likely to arise in late processing stages. However, this only holds for exemplar effects of indexical variation, since the opposite pattern (exemplar effects in early processing stages) is predicted for allophonic variation.

Under the assumption that exemplars are part of episodic memory, we formulated four predictions on when exemplar effects in auditory identity priming experiments should be largest. We then reviewed existing literature to verify to what extent the results of previously reported experiments are in accordance with these predictions. Predictions I, II and III are borne out in our review of the relevant literature. Prediction I states that exemplar effects are largest when participants must use their episodic memories because of task requirements. This prediction was supported by a single study by Luce and Lyons (1998). Some incongruent evidence was also reported (exemplar effects arose for tasks which do not require participants' use of episodic memory, e.g., Church & Schacter, 1994); these findings are probably due to the use of degraded stimuli, which may have stimulated participants' reliance on episodic memory (see Prediction II).

Prediction II states that larger exemplar effects arise for less intelligible stimuli, since adverse listening conditions may invite participants to use their episodic memories in identity priming tasks. This prediction was supported by Mattys and Liss (2008) and Chapter 4 of this dissertation, amongst others. These studies reported larger exemplar effects for stimuli produced by dysarthric speakers (Mattys & Liss, 2008) or stimuli embedded in noise (Chapter 4) than for clear stimuli.

It should be noted that this prediction concerns target stimuli (presented to participants in the test phase, i.e., at retrieval). For primes, a different outcome is expected: encoding may suffer from stimulus degradation because part of listeners' attentional resources must be allocated to stimulus disambiguation (the effortfulness hypothesis, Rabbitt, 1968). Exemplar effects may therefore decrease if primes are less intelligible. No study has tested this; intelligibility manipulations were usually applied across the board (i.e., to both primes and targets) or to targets only (as part of some other manipulation; e.g., Church & Schacter, 1994; Goldinger, 1996; Schacter & Church, 1992; Theodore et al., 2015). Importantly, one of these studies (reported in Chapter 4 of this dissertation) provides initial evidence for a role of prime intelligibility. Exemplar effects in this study only arose when participants processed primes fast, which reflected easy encoding. Further investigation is, however, necessary to establish how prime intelligibility precisely affects the occurrence of exemplar effects.

Prediction III states that exemplar effects arise mostly for materials that are not consistent with prior knowledge. This prediction was confirmed by several studies: Goldinger (1996, 1998), Dufour et al. (2017) and Dufour and Nguyen (2014) reported larger exemplar effects for uncommon word types (i.e., with low frequencies of occurrence), and Nygaard et al. (2000) found exemplar effects only for stimuli produced in unusual ways (i.e., at a slow speech rate, with loud vocal effort, or with high amplitude).

Prediction IV states that exemplar effects are larger when participants engage in perceptual rather than meaning-related processing of primes. Two studies provided evidence in line with this prediction (Goldinger, 1996; Theodore et al., 2015). However, there may be confounds in these studies. For instance, the perceptual tasks required participants to focus on the acoustic information that formed the basis for the prime-target match versus mismatch condition whereas the meaning-related tasks did not.

The literature thus provides experimental results that are in line with our main hypothesis stating that exemplars reside in episodic memory. It is, however, also clear that more research is necessary. In this direction, we conducted two new experiments.

In these experiments, we investigated two of the four predictions more directly (Predictions I and IV). These state that exemplar effects are largest when participants are required to use their episodic memories because of the experimental task during test (Prediction I) and when participants focus on perceptual rather than meaning-related properties of primes during familiarization (Prediction IV). We tested these predictions by implementing different tasks in the experiments' familiarization and test phases. Primes and targets were spoken by a male or female speaker, and prime-target pairs were presented in the same or a different voice; we tested whether targets were recognized more quickly and/or more accurately when they were spoken in the same voice as their primes.

In Experiment 1, participants engaged in old-new judgment during test, a task that requires the use of episodic memory. In Experiment 2, participants engaged in semantic classification during test (animacy judgments). This task does not require participants' use of episodic memory. This allowed us to test the prediction that exemplar effects are larger when participants use their episodic memories.

To test whether exemplar effects are larger when participants focus on primes' perceptual properties, participants in both experiments engaged in one of three tasks during the familiarization phase: loudness judgment, voice judgment, or semantic classification (indoor/outdoor judgments). The loudness and voice judgment tasks direct participants' attention to perceptual aspects of the primes, while the semantic task directs participants' attention to primes' meanings. We used two perceptual tasks to also assess whether exemplar effects are larger when participants pay attention to the acoustic properties which form the basis of the prime-target match versus mismatch condition. Paying attention to the acoustic properties of the match/mismatch condition was necessary in the voice task, but not in the loudness task.

We also instructed participants in all three familiarization tasks in both experiments to pay attention to words identity because that would be important for later in the experiment. All though it may be argued that this instruction invites an 'episodic mode' in both experiments, and thereby makes Experiment 2 resemble Experiment 1 to a certain extent, this specific instruction was necessary because of our experimental set-up. We needed to divide the experiment into two parts (both consisting of a familiarization and a test phase) to ensure the lag between primes and targets remained small. If we would not have instructed participants to pay attention to word identity, their behavior in the second familiarization phase (i.e., after they knew they would be questioned about the words in the familiarization phase) would be very different from their behavior in the first familiarization phase. In order to keep participants' behavior in the familiarization phases comparable between the experiments, we added a visual old-new task after each test phase in Experiment 2.

Our experiments differed from those reported in published papers in a number of ways. First, while manipulating tasks, we ensured all other aspects were identical between experiments (i.e., we only altered task instructions). This critically improved the comparison of the effects of tasks. Second, almost all previous studies used identical tokens for primes and targets in the match condition. Instead, we always used different tokens for primes and targets because this enhances the ecological validity of our findings. Finally, we took differences in stimulus intelligibility and prime-token similarity between our speakers into account by post-hoc exploration of differences in exemplar effects between our two speakers.

We established exemplar effects in Experiment 1 in the RT data, which held for both speakers. In this experiment, participants used their episodic memories to perform the old-new judgment task. This result is in line with other studies that obtained exemplar effects when participants needed to use their episodic memories (e.g., Luce & Lyons, 1998). Moreover, this finding offers support with the main hypothesis of this study, namely that exemplars are represented in episodic memory.

We obtained exemplar effects even though we used non-identical tokens for primes and targets. This finding adds to the few articles that show that exemplar effects may also arise when more realistic stimulus materials are used (e.g., Hanique et al., 2013), and therefore shows that exemplar effects are not merely a by-product of the repetition of identical tokens.

The exemplar effects were not modulated by the type of task that participants engaged in during the familiarization phase. Hence, we did not find support for Prediction IV, stating that exemplar effects are larger when participants pay attention to perceptual properties of primes. Our results also do not indicate that exemplar effects are larger when, during familiarization, participants pay attention to the acoustic information that forms the basis of the prime-target match versus mismatch condition (i.e., in the voice classification task). The lack of an effect of familiarization task is in line with other null results (e.g., Scheffert, 1998; Naveh-Benjamin & Craik, 1995), but is at odds with Goldinger (1996) and Theodore et al. (2015), who established larger exemplar effects for perceptual tasks.

Possibly, this discrepancy with Goldinger (1996) and Theodore et al. (2015) stems from differences between the exact familiarization tasks that were compared; for instance, both Goldinger (1996) and Theodore et al. (2015) used metalinguistic tasks (i.e., initial phoneme classification and syntactic category classification), while our tasks (loudness, voice or indoor/outdoor judgments) depend less on participants' use of metalinguistic knowledge. Another possible explanation for our null result is a lack of statistical power in our study. For instance, Goldinger tested 35 participants on 150 stimulus words, whereas we tested 31 participants on no more than 36 words per familiarization task. Assuming similar variance in the data of the two studies, Goldinger could therefore detect smaller effects than we could.

Our post-hoc analyses, testing for interactions with speaker, suggest that the exemplar effects were robust over speakers in the RT data. In the accuracy data, however, where they were only present for the male speaker. A priori, we did not expect to see differences between our speakers, given that they were both normal, healthy talkers. In addition, previous studies with multiple talkers provide no indication that exemplar effects may be speaker-dependent (Hanique et al., 2013, Chapter 4 of this dissertation). This raises the question what could have caused the difference in exemplar effects between our speakers in the accuracy data.

Our rating experiments showed that – even though both speakers could be understood well – the male speaker was perceived as more intelligible than the female speaker. We also found that the male speaker's prime-target tokens sounded more similar to each other than those of the female speaker. We deem it unlikely that the relatively high intelligibility of the male speaker was responsible for the difference in exemplar effects between speakers we found. First, although the male speaker was more intelligible than the female speaker, the female speaker was intelligible as well. Second, previous studies showed larger rather than smaller exemplar effects for less intelligible materials (e.g., Mattys & Liss, 2008, see also Prediction II).

Instead, the higher prime-target token similarity of the male speaker probably caused the difference in exemplar effects between speakers that appeared to be present in Experiment 1. If this is true, this would imply that when non-identical tokens are used in the match condition, it is not enough to be uttered by the same speaker. Further research into speaker differences is necessary to establish how speaker idiosyncrasies interact with exemplar effects. Note that this result is telling for the question to what extent exemplar effects generalize to more realistic listening circumstances. Even the natural item-to-item pronunciation variation that a normal speaker exhibits when carefully producing isolated word tokens of which only good ones were selected for inclusion in the experiment may be enough to prevent exemplar effects from arising. This result, together with previously established restrictions on the occurrence of exemplar effects (see e.g., Nygaard et al., 2000), therefore posits severe limitations on the robustness of exemplar effects, and hints at a small role for exemplars in everyday listening circumstances.

In Experiment 2, our main analyses showed null results in both response accuracy and RTs. Hence, no exemplar effects arose when participants performed a semantic categorization task at test. This results contrasts sharply with Experiment 1, in which participants performed old-new judgment, and where exemplar effects arose. In a combined analysis of the RTs of Experiments 1 and 2, we obtained an interaction which confirmed that exemplar effects were only present in Experiment 1. This difference between Experiments 1 and 2 is in line with Prediction I stating that exemplar effects arise especially when participants use their episodic memories.

Like in Experiment 1, post-hoc analyses suggest differences in exemplar effects between our speakers in Experiment 2. In response accuracy (data which should, however, be interpreted with caution due to performance at ceiling), the occurrence of exemplar effects was restricted to targets produced by the male speaker, as was the case in Experiment 1. For the female speaker, we observed a reversed effect in this experiment (while in Experiment 1, we obtained a null result for this speaker).

We wonder whether the pattern in the accuracy results should have the same explanation in both experiments, given that the experiments show very different patterns in the reaction times. In the RTs of Experiment 2, exemplar effects seem present for the targets produced by the male speaker, while a reversed effect appeared to arise for the targets produced by the female speaker (faster responses to targets spoken in the other voice as their primes). Experiment 1 did not show such a reversed effect; neither does existing literature. Exemplar theory, which predicts beneficial effects of acoustic similarity on word recognition, cannot easily account for this finding: two tokens by the same speaker should always be more similar than two tokens produced by two highly distinct speakers. This raises the question what caused this reversed effect.

Importantly, the post-hoc results of Experiment 2 can be summarized as participants being faster if primes were uttered by the male speaker, that is, the more intelligible speaker. Since this facilitation arose independently of the speaker of the target, it must reflect abstract priming rather than an exemplar effect. If a prime is clearly perceivable, it activates its word form well. When the prime word is subsequently repeated as a target, its word form is still active and thus facilitates the recognition of the target. Conversely, less intelligible primes (such as those uttered by our female speaker) may activate their word forms less. The recognition of targets following such primes is therefore not expected to be strongly facilitated. Hence, the facilitation for targets produced by the female speaker in the mismatch condition was likely due to the high intelligibility of the primes produced by the male speaker. If this is true, the fast responses to targets produced by the male speaker in the match condition were probably also caused by good prime intelligibility.

Like the main analyses, the post-hoc analyses in Experiment 2 this did not show clear exemplar effects. Rather, they suggest general facilitation due to the highly intelligible primes. This again shows that Experiment 2, in which participants performed a task for which their use of episodic memory was not necessary, yielded different results than Experiment 1, in which participants needed to use their episodic memories. As such, the results of the post-hoc analyses of the two experiments are also in accordance with our hypothesis that exemplars are part of episodic memory (Prediction I).

A more general lesson to be drawn from the post-hoc analyses in Experiment 2 is that findings that appear to be exemplar effects (such as the effect we obtained for targets produced by our male speaker) may in fact be the product of another mechanism. It was only through the reversed effect for targets produced by our female speaker that we learned that the targets of both our speakers probably just showed abstract priming. This shows that we should be careful in concluding that effects are caused by the activation of exemplars, given that general facilitation due to favorable encoding circumstances may produce similar outcomes. It calls for replication of previous studies with stimuli from different speakers.

We found an effect of lag between prime and target in Experiment 2, indicating that participants recognized targets more quickly if they were further away from their primes. This result probably reflects that participants were aware of repetitions in the experiment: they probably noticed that all primes reappeared as targets at some point during the test phase. As a consequence, the further targets were away from their primes, the greater their likelihood.

In conclusion, we have three key findings. First, we found that existing experimental results on exemplar effects are in line with the hypothesis that exemplars are part of episodic memory. Second, in our experiments, we established exemplar effects when participants had to use their episodic memories. When they carried out semantic classification, a task that does not require the use of episodic memory, participants appeared to rely on abstract lexical representations. Third, our data suggest that there are limits on the amount of item-to-item variability against which exemplar effects are robust, since the effects in response accuracy in Experiment 1 disappeared for the speaker whose primes and targets sounded less similar. Together, our results are in line with the hypothesis that exemplars are part of episodic memory, and suggest a hybrid model in which episodic exemplar-based representations exist alongside abstract lexical representations.

ERPs but not behavioral measures show that exemplar effects differ depending on whether participants use their episodic memories

Chapter 3

This chapter is a reformatted version of

Annika Nijveld, Kimberley Mulder, Louis ten Bosch and Mirjam Ernestus (submitted). ERPs but not behavioral measures show that exemplar effects differ depending on whether participants use their episodic memories.

Abstract

The nature and role of exemplars in speech comprehension appear from the circumstances under which exemplar effects arise. A number of previous studies have investigated which experimental tasks lead to exemplar effects, but the results of these studies as a whole are inconclusive. We argue that some of the reported null results may be due to the use of relatively insensitive experimental measures (response accuracy and reaction times). To investigate the emergence of exemplar effects in more detail, we compared two experimental tasks (old-new judgment and semantic classification) measuring participants' response accuracy, reaction times, and EEG signals. The behavioral data showed exemplar effects for neither task. In contrast, the ERPs showed clear exemplar effects in the earliest time windows for the old-new judgment task. In the semantic classification task, we observed a reversed effect, which may indicate that exemplar effects are manifested differently depending on task. Alternatively, this mismatch effect may stem from participants relying on abstract representations. Our combined behavioral and EEG results show that behavioral measures may often be too insensitive to detect exemplar effects, which may explain the null results in many studies addressing exemplar effects. Moreover, our results for the two tasks support the hypothesis that exemplar effects vary specifically depending on whether listeners rely on their episodic memories.

Introduction

Theories of speech comprehension distinguish two types of representations for the pronunciation of words: abstract representations and clouds of exemplars. An abstract representation mostly consists of a single phonological representation for a word, and does not contain details about specific tokens of the word (such as indexical and prosodic information about the speaker's gender, age, and mood). Exemplars, in contrast, mentally represent each occurrence of a word in full phonetic detail, and all processed occurrences of a word together form a cloud of exemplars for that word.

The assumption that word tokens are (at least temporarily) stored in full phonetic detail is supported by numerous auditory identity priming experiments. In these studies, spoken words were repeated once, and listeners were quicker and/or more accurate to recognize word repetitions if the two occurrences of a word (the 'prime' and 'target' tokens) shared perceptual properties such as the speaker's voice, the realization of a certain phoneme, or the speech rate (the 'match' condition) than when the two occurrences did not share such properties (the 'mismatch' condition; e.g., Bradlow et al., 1999; Craik & Kirsner, 1974; McLennan et al., 2003; Strori et al., 2018). The assumption is that listeners keep the acoustic-phonetic details of the prime in memory in the form of an exemplar, which modulates the later recognition of the target.

Exemplar effects did not arise in all priming experiments. Importantly, exemplar effects often arose in experiments using tasks that required participants to make explicit memory judgments, such as a task in which participants have to judge whether words occurred previously in the experiment or not (an old-new judgment task; e.g., Bradlow et al., 1999; Goh, 2005; Saldaña et al., 1996; Goldinger, 1996), while the effects often failed to arise for tasks which lacked such an explicit memory component, such as lexical decision (e.g., Hanique et al., 2013; McLennan et al., 2003; McLennan & Luce, 2005, and Chapter 2 of this dissertation).

A limited number of previous studies investigated the effect of task directly. As expected, one study found that exemplar effects only arose for an old-new judgment task (a task which requires explicit episodic memory judgments), but not for a lexical decision task on the same stimuli (a task which does not require explicit episodic memory judgments; Luce & Lyons, 1998). However, other studies did not observe that exemplar effects arise most clearly when participants are explicitly required to rely on their episodic memories. Goldinger (1996) and Pilotti, Bergman, Gallo, Sommers, and Roediger-III (2000) obtained exemplar effects both for tasks which do and do not require participants' explicit reliance on episodic memory, and Schacter and Church (1992) and Church and Schacter (1994) found that exemplar effects arose mostly for tasks that do not require explicit memory judgments. These results may be explained by the experimental design chosen: task manipulations often not only included of different instructions to participants, but also the use of different stimulus materials between tasks. For instance, stimuli in one but not the other type of task were degraded by low-pass filtering, or were substantially masked by background noise. Previous work has shown that specific characteristics of the stimuli used in an experiment (e.g., their frequency of occurrence, or their signal-to-noise ratio) may also influence the occurrence of exemplar effects (e.g., Dufour & Nguyen, 2014, and Chapter 4 of this dissertation). As a result of the varying experimental designs, these studies could thus not clearly differentiate which tasks favor exemplar effects.

Exactly which tasks most reliably evoke exemplar effects is informative about the nature and role of exemplars. If exemplar effects only consistently arise when participants need to make explicit memory judgments, exemplars may not be represented in the mental lexicon (as assumed for instance by the exemplar model of Goldinger, 1998), but instead in episodic memory. Importantly, this would indicate that the role of exemplars in speech comprehension is much more limited than has previously been assumed.

We investigated further whether exemplar effects arise more reliably in a task that requires explicit episodic memory judgments versus a task that does not. In our study, one group of participants engaged in an old-new judgment task, while another group of participants engaged in a semantic classification task, in which they judged whether words had animate or inanimate referents. Importantly, only the old-new judgment task requires participants' explicit memory judgments on recent input. We used a semantic classification task rather than lexical decision because this allowed us to use exactly the same real word stimuli in both tasks. Primes and targets in our study were presented in the same or a different voice.

One reason exemplar effects do not consistently arise might be that exemplar effects are typically small in effect size (e.g., the effects often range in tens of milliseconds in participants' reaction times). As such, it is possible that the effects were present in some of the experiments that produced null results, but were too weak to be detected. We hypothesize that the dependent measure mostly used in the literature, i.e., participants' response accuracy and sometimes their reaction times (RTs), may not have been sensitive enough to capture exemplar effects in all cases. Therefore, exemplar effects may better be captured with more sensitive methods than behavioral methods.

One more sensitive method is eye-tracking, a non-invasive method that monitors participants' gaze fixations with millisecond-by-millisecond resolution as the speech signal unfolds over time. As this method allows for an on-line measurement of behavior, it is considered a more fine-grained measure of cognitive processes than, for instance, key presses. In addition, while overt behavioral responses are the outcome of an accumulation of processes, including motor processes, eye-tracking taps into a single processing stage. Papesh, Goldinger, and Hout (2016) used this method to examine the occurrence of exemplar effects. This study found that exemplar effects were especially evident in the two eye-tracking measures (i.e., in eye movement initiation and fixation times), while surfacing less clearly in participants' overt behavior (i.e., click RTs). Exemplar effects arose most clearly in the earliest eye-tracking measure (i.e., eye movement initiation times).

Another more sensitive method is EEG (electroencephalography). EEG-based methods are also considered to offer a more process-pure estimate of perception than behavioral measures, as these methods, like eye-tracking, register on-line processes more directly than behavioral overt responses do. An important advantage of EEG over eye-tracking is that ERPs (event-related potentials: stimulus-locked brain potentials derived from the EEG signal) are manifested in different components with specific neural signatures in timing, polarity, and scalp distribution. Importantly, such components are functionally associated with distinct stages of spoken word recognition. This allows for conclusions about which stages of processing are affected by experimental manipulations. For instance, the N400 component (a negative, central-parietal component peaking around 400 ms after word onset) has been linked to lexical processing, going from activation of a set of word candidates to the selection of the target word (e.g., Desroches, Newman, & Joanisse, 2009). In our study, we measured participants' response accuracy, their RTs as well as their EEG.

A number of studies in the auditory or cross-modal domain have successfully applied ERPs to study exemplars (Campeanu, Craik, Backer, & Alain, 2014; Dufour et al., 2017; Friedrich, Kotz, Friederici, & Alter, 2004; Schild, Becker, & Friedrich, 2014a, 2014b). In most of these studies, EEG was administered in addition to behavioral measures. ERP exemplar effects arose in different time windows, and with varying scalp topographies. Most of these ERP exemplar effects were linked to specific ERP components.

In the earlier time windows (ranging from approximately 100 to 400 ms after word onset), exemplar effects arose over posterior electrode sites (Friedrich, Kotz, Friederici, & Alter, 2004; Schild et al., 2014b). Schild et al. (2014b) tested for exemplar effects of stress (in a design with spoken initial syllables as primes and spoken full words as targets). Schild and colleagues observed more negative amplitude deflections for the match than the mismatch condition in the 100-300 ms window. This time window is associated with early acoustic processing, as indexed by the N100 and P200 components. The authors did not specifically link their effects in the 100-300 ms window to these ERP components, however, as their combined effects ranged over a larger time window (which they linked to a central N400-like negativity). Friedrich, Kotz, Friederici, and Alter (2004) tested for exemplar effects of stress in a cross-modal design, in which spoken initial syllables served as primes, and written words served as targets. This study observed more negative ERP effects for the match than the mismatch condition in the 300-400 window. Friedrich and colleagues linked these exemplar effects to the P350, a positive component peaking around 350 ms after word onset, which is associated with lexical identification (a component first described in Friedrich, Kotz, Friederici, & Gunter, 2004). The reduction of this P350 component in the match condition was interpreted as reflecting facilitated lexical identification.

In the later time windows (ranging from approximately 400 to 1000 ms after word onset), most studies tested for modulations of the N400 component. The N400 is a frequently observed ERP component in different language-related tasks (see Kutas & Federmeier, 2011, for a review). This component is modulated by a range of factors that affect lexical access. Relevant to our purposes, it is sensitive to priming and stimulus congruency: its amplitude is reduced for congruent targets (e.g., Kutas & Van Petten, 1988). Compatibly, Schild et al. (2014a) and Schild et al. (2014b) observed reduced N400s for spoken targets which matched spoken prime syllables in stress (over anterior and posterior brain regions, respectively). Across and within studies, however, this negative polarity was not entirely stable, as Dufour et al. (2017) noted a *positive* deflection for a match, as did Schild et al. (2014a) and Schild et al. (2014b) for other brain regions tested in their studies (i.e., posterior and anterior sites, respectively). The study by Dufour and colleagues tested for exemplar effects of speaker voice, like our study. Furthermore, one other study, conducted by Campeanu et al. (2014), observed more negative amplitude deflections for a match than a mismatch in speaker voice over anterior electrodes across a fairly large time window, ranging from 200 to 650 ms relative to word onset. The authors functionally linked these effects to familiarity, one of the components of recognition memory, which reflects a global measure of memory strength or stimulus recency (e.g., Yonelinas, 2002).

Except for Dufour et al. (2017), the ERP studies discussed above collected behavioral measurements alongside participants' EEG. Importantly, the exemplar effects that surfaced in the ERPs were not fully replicated in participants' behavior in any of these studies. Schild et al. (2014a) collected RTs next to ERPs, in which exemplar effects did not surface at all. Friedrich, Kotz, Friederici, and Alter (2004) collected RTs and accuracy alongside ERPs, and established exemplar effects in the RTs, but not in accuracy. In Campeanu et al. (2014) and Schild et al. (2014b), exemplar effects arose in the behavioral measures (RTs in Schild et al., 2014b, and RTs and accuracy in Campeanu et al., 2014); these exemplar effects, however, were restricted to subsets of the experimental conditions (and did thus not arise across the board in participants' behavior as they did in the ERPs). These results again hint that ERPs are more sensitive to detect exemplar effects than behavioral measures are.

Another cause for exemplar effects not consistently surfacing in previous studies might have been the way data were statistically analyzed. In most of the behavioral and neurophysiological studies that investigated exemplar effects, the most recent statistical techniques now used in the field were not yet available. Over the last years, statistical methods such as linear mixed effects regression modelling (e.g., Pinheiro & Bates, 2000) have substantially been improved. Linear mixed effects regression techniques offer an array of advantages compared to traditional ANOVAs (Analysis of Variance), including the possibility to let participants and items vary in their sensitivity to experimental effects (i.e., by modelling of random slopes), and the better conservation of statistical power in the presence of missing observations. We used linear mixed effects regression modelling to more precisely model the contribution of exemplar matching to our dependent variables (RTs, accuracy, and EEG) than the older traditional statistical methods would have allowed for.

In summary, the tasks which favor exemplar effects are informative about the nature of exemplars, and about exemplars' role in speech comprehension. In this study, we investigate to what extent the experimental task participants engage in affects the size or occurrence of exemplar effects. Specifically, our research question is whether exemplar effects surface more reliably in tasks that are explicitly memory-based versus tasks that are not. As a secondary research question, we investigate whether task-dependent exemplars effects are more adequately measured with EEG than with behavioral measures.

To address our research questions, we conducted a long-term auditory repetition priming experiment. This experiment comprised of two parts, each consisting of a familiarization and a test phase and each presenting half of the stimuli (see Table 3.1). Primes were presented during the familiarization phase, and targets during the test phase. During the familiarization phases, all participants engaged in a loudness judgment task. This task focuses participants' attention on the perceptual aspects of the speech signal. A focus on words' perceptual aspects (rather than, for instance, their meaning-related properties) during familiarization has been shown to enhance the likelihood of obtaining exemplar effects (e.g., Goldinger, 1996; Theodore et al., 2015). We used a loudness-based rather than a speaker voice-based judgment task because we wished to not specifically alert participants of the variation that formed the basis of the match/mismatch condition in the experiment. During the test phases, one group of participants engaged in an old-new judgment task, and one group of participants engaged in a semantic classification task. For each task in the test phases, we tested whether participants' processing differed for targets in the same versus the other voice as their primes by measuring participants' RTs, accuracy and EEG.

Method

Participants

Sixty-six right-handed Dutch native listeners participated in the experiment. Of these participants, 33 were randomly assigned to the old-new judgment task (9 male, 18-30 years old, mean: 22 years old), and 33 to the semantic classification task (6 male, 18-30 years old, mean: 22 years old). All were highly educated, were paid for their participation, reported good hearing, and had no known neurological disorders. All participants gave their written informed consent.

Materials

The experiment contained 64 experimental real words, 64 real word fillers, and four real word practice items. The experimental words were Dutch bisyllabic nouns with inanimate referents (e.g., *tafel* 'table', see also the Appendix), whose log-transformed frequencies of occurrence ranged between 0.3 and 4.3 (average: 2.2; *SD* = 0.8; counts of the SUBTLEX-NL corpus, Keuleers, Brysbaert, & New, 2010). The 64 fillers were bisyllabic nouns with animate referents (e.g., *oma* 'grandmother'), and were matched to the experimental words in log-frequency of occurrence (range: 0.3 - 4.5; on average 2.2; *SD* = 0.8, which is not statistically different from the frequencies of occurrence of the experimental words: *t* (63) = 0.17, *p* = .09). Of the four practice items, two were animate and two were inanimate bisyllabic nouns (average log-frequency occurrence: 2.4; *SD* = 0.1; range: 2.3 - 2.5).

As mentioned above, the experiment consisted of two parts, each consisting of a familiarization and a test phase. During each familiarization phase, we presented 32 experimental words as primes. During each test phase, we presented the 32 experimental words from the familiarization phase as targets, in addition to 32 fillers. Each familiarization phase started with two practice items, which we repeated at the start of the following test phase. This design, summarized in Table 3.1, yields a total

Table 3.1: Overview of tasks and stimuli used in the familiarization and test phases of the two parts of the experiment. Of each stimulus category, an example is given. The participants engaging in the semantic classification task during the test phases performed an additional old-new judgment task on paper after each test phase which is not listed in this table. ‘Old-new’ denotes old-new judgment, and ‘sem. classification’ ‘semantic classification’.

Part 1			
Familiarization phase 1 task: loudness judgment		**Test phase 1** task: old-new *or* sem. classification	
2 practice words	*strijder* ‘warrior’	2 practice words	*strijder* ‘warrior’
32 primes	*tafel* ‘table’	32 targets	*tafel* ‘table’
		32 fillers	*oma* ‘grandmother’
Part 2			
Familiarization phase 2 task: loudness judgment		**Test phase 2** task: old-new *or* sem. classification	
2 practice words	*donut* ‘doughnut’	2 practice words	*donut* ‘doughnut’
32 primes	*anker* ‘anchor’	32 targets	*anker* ‘anchor’
		32 fillers	*vlinder* ‘butterfly’

of 200 trials (i.e., 100 per part). The division of the experiment into two parts served to keep primes and targets close together, as previous research has shown that a separation of primes and targets by too many intervening trials (i.e., > 100 trials) diminishes the chances of finding exemplar effects (Hanique et al., 2013).

Half of the primes were presented in the voice of a male speaker, and half in the voice of a female speaker, both native speakers of Dutch. All targets were presented in the voice of the male speaker. The speaker match condition meant that prime and target were both uttered by the male speaker, while the speaker mismatch condition meant that the prime was uttered by the female speaker, and the target by the male speaker (see also Table 3.2). Primes and targets were always different tokens. To not stand out from the targets, fillers (occurring in the test phases alongside the targets) were also all uttered by the male speaker. Practice items at the start of the familiarization phases were uttered by the male or the female speaker, and those at the start of the test phases by the male speaker.

Table 3.2: Experimental conditions and examples of primes and targets in the experiment.

Condition	Prime	Target
Speaker match		
male speaker - male speaker	*tafel*$_{\text{male}}$ ‘table’	*tafel*$_{\text{male}}$ ‘table’
Speaker mismatch		
female speaker - male speaker	*anker*$_{\text{female}}$ ‘anchor’	*anker*$_{\text{male}}$ ‘anchor’

For the primes and practice items, we recorded tokens with the male and the female speaker in a sound-attenuating booth at a 44.1 kHz sampling rate (16 bits/sample). We selected one token from each speaker from the recordings for the primes, and two tokens from each speaker from recordings for the practice items. Targets and fillers were recorded with the male speaker only, and one token was selected for each word. The prime and target tokens we selected from the male speaker (N = 128) had an average word duration of 589 ms (*SD* = 113; range: 330 - 831 ms), while the prime tokens produced by the female speaker (N = 64) were 618 ms long on average (*SD* = 93; range: 428 - 842 ms), which is not different statistically according to a linear model ($\hat{\beta}_{\text{female speaker}} = 28.4$, $t = 1.7$, $p = .08$)

Even though all stimuli were equalized at 70 dB, the equalized stimuli are likely to differ in their perceived loudness. This is the case because the dB scale is not corrected for the (non-linear) frequency-dependent sensitivity of the human hearing system. In addition, the 70 dB is the average amplitude across the signal, and individual stimulus tokens differ in the way energy is distributed along the signal. Editing (e.g., cutting of the long audio file into individual stimulus wave files) was done in Praat (Boersma & Weenink, 2018).

We created three lists of all stimuli (apart from the practice items) with different word orders, with three constraints: no more than three consecutive primes in the same speaker voice occurred, no more than three consecutive targets or fillers occurred, and primes and targets were separated by no more than 100 trials (since longer separations diminish the chances of finding exemplar effects, Hanique et al., 2013). From these lists, we derived three more lists by swapping the two parts of the experiment, resulting in a new total of six lists. These six lists were then used to create six additional lists which had the same word order, but with the prime spoken by the other speaker. The eventual total of lists was twelve. Each participant was presented with one list.

Procedure

We tested participants individually in a sound-attenuating booth, and presented auditory stimuli via closed headphones at a comfortable listening level using Presentation software (Version 16.4, Neurobehavioral Systems, Inc., Berkeley, CA, www.neurobs.com). For the familiarization phases, we instructed all participants to indicate whether they thought the prime presented in the current trial sounded equally loud as the prime in the previous trial, or not ('equally loud'-responses with a press on a button box with the dominant right hand; also see Table 3.1 for an overview of all of the tasks used in experiment). On the first trial, participants were instructed not to make a response but to just listen. In addition to performing loudness judgments, participants were asked to pay attention to the words themselves to ensure encoding of word identity.

For the test phases, one group of participants was instructed to indicate as quickly and accurately as possible whether or not the word they heard had occurred previously in the experiment ('old'-responses with the dominant right hand). The other group of participants was instructed to indicate as quickly and accurately as possible whether the word they heard referred to a living or a non-living object ('non-living'-responses with the dominant right hand). This latter group of participants also carried out an old-new judgment task on paper after each test phase in order to equally reward them for having paid attention to word identity during the familiarization phase as the old-new judgment participant group. After the first part of the experiment (i.e., after one familiarization phase and one test phase, and for the semantic classification group one old-new judgment task on paper), both groups of participants performed a task in which they solved arithmetic problems on paper for four minutes to clear their memory to some extent and to let some time pass between the two parts of the experiment.

A trial consisted of an empty grey screen (for 500 ms), a grey screen with a white fixation cross (30x2 by 30x2 pixels) in the centre (for 1000 ms), an empty grey screen (for 500 ms) followed by the auditory stimulus, presented while the screen was grey. After word onset, the next trial was initiated at the participant's response or after a time-out of 3500 ms. The experiment lasted approximately fifteen minutes.

EEG recording

We collected EEG data from 64 Ag-Ag Cl active electrodes positioned according to the 10-20 standard system in an elastic cap (actiCAP system, Brain Products GmbH, Munich, Germany; 8 midline electrodes and 50 lateral electrodes). Bipolar and vertical electrooculograms (EOG) were recorded to monitor for blinks and saccades (using four electrodes). We used the left mastoid as online reference electrode, and an additional electrode was placed on the participant's right mastoid for re-referencing offline. We kept electrode impedances below 50 kΩ. The EEG was recorded continuously with a 0.02 - 100 Hz band-pass filter, and digitized with a 500 Hz sampling frequency.

Behavioral analyses

We analyzed participants' response accuracy to experimental words in the test phases (i.e., the targets) by means of generalized mixed effects regression models with the binomial link function, and we analyzed log-transformed reaction times (log RTs; measured from word onset) with linear mixed effects regression models (with the `lme4` package, Bates et al., 2015). We logarithmically transformed RTs to reduce skewness in their distribution (e.g., Baayen & Milin, 2010).

We run all statistical tests in `R` statistical software (RCoreTeam, 2018), and we tested for the interaction between our predictors of interest *Speaker match condition* (which reflected whether a prime and target were spoken by the same speaker or not) and *Task* (old-new judgment or semantic classification). In addition, we tested for effects of a number of control predictors known to affect speech processing: log-transformed *Word duration*, log-transformed *Word frequency* as obtained from the SUBTLEX-NL corpus (Keuleers et al., 2010), *Trial number*, *Part* of the experiment (part 1 or 2), *Lag* in trials between prime and target, and, for the log RT analyses, log-transformed reaction times to the prime (*RT prime*) and to the preceding stimulus (*RT preceding*). We only included significant fixed effects and interactions in the models (i.e., whose absolute t-values exceeded 1.96, which implies p < .05, as the data approaches a normal distribution given its many observations), as well as simple effects of predictors appearing in significant interactions.

We used *Word* and *Participant* as crossed random effects (intercepts), and tested for random slopes for the predictors of interest and their interactions (*Task*, *Speaker match condition* and their interaction by-word, and *Speaker match condition* by-participant). We only included random slopes if they did not lead to model overfitting, as evidenced by model convergence issues (i.e., warning or errors), or high (> 0.6) correlations between the random intercepts and slopes. Inclusion of these random slopes did not result in significant improvements of the statistical models in any of our behavioral analyses (assessed with chi-square tests of nested models with the `anova()` function of the `car` package; Fox & Weisberg, 2011).

We did not test for random slopes of control predictors for three reasons: first, we had no experimental hypotheses about those; second, doing so increases the chance of overfitting the models to this particular dataset (which would decrease the generalizability of our findings); and third, doing so increases the chances of model convergence failures. We restricted the log RT analyses to targets which received correct responses, and to trials whose log RTs were within two standard deviations from the grand mean. For the best log RT model, we removed data points with standardized residuals exceeding 2.5 standard deviation units and refitted the model.

ERP analysis

We used FieldTrip (Oostenveld, Fries, Maris, & Schoffelen, 2011) for preprocessing the raw EEG data. We first re-referenced the EEG data offline to the average of the left and right mastoids, and filtered the data with a high cut-off filter of 35 Hz. We then segmented the data into epochs from -200 to 1000 ms relative to stimulus onset.

We manually rejected noisy trials, and subsequently applied Independent Component Analysis (ICA using the ‘runica’ application) to remove ocular artefacts from the data. Noisy electrodes were discarded before ICA and interpolated with the average of neighbouring electrodes after correction by ICA. We further cleaned the data by ensuring that all traces remained within the 2.5 *SD* band around the mean of all traces (leaving ±90% of all traces).

In line with previously reported ERP exemplar effects (e.g., Friedrich, Kotz, Friederici, & Alter, 2004; Schild et al., 2014b; Dufour et al., 2017), we focused our analyses on three time windows (illustrated in Figure 3.2): the auditory N100/P200 window (100-300 ms, e.g. Schild et al., 2014b), the P350 window (300-400 ms, e.g., Friedrich, Kotz, Friederici, & Alter, 2004; Schild et al., 2014b), and the extended N400 window (400-800 ms, e.g., Dufour et al., 2017; Schild et al., 2014b). The exemplar effects in the literature in the two earlier windows were mostly found on posterior sites; we therefore restricted the analyses in these windows to the posterior electrodes (i.e., 34 electrodes: T7, C5, C3, C1, Cz, C2, C4, C6, T8, TP7, CP5, CP3, CP1, CPz, CP2, CP4, CP6, TP8, P7, P5, P3, P1, Pz, P2, P4, P6, P8, PO7, PO3, POz, PO4, PO8, O1, and O2). The N400 typically peaks more widely distributed topographically; we therefore applied the analysis in this time window to all 58 electrodes.

As in the behavioral analyses, we tested for the interaction between *Speaker match condition* and *Task*. We calculated mean amplitudes per trial, electrode (for the 34 posterior electrodes, or all 58 electrodes), speaker match condition (match or mismatch), and participant for each time window, and submitted these to linear mixed effects regression models. Since we had no further predictions about the effects' topographical distributions, we did not test for interactions with electrode site.

For compatibility with related previous ERP work (e.g. Campeanu et al., 2014; Dufour et al., 2017), we also report ANOVAs (computed with the `ez` package; Lawrence, 2016). In contrast to various recent ERP studies, we did not use cluster-based permutation tests (e.g., Maris & Oostenveld, 2007), as these are not suitable to test for interactions that involve different participant groups (since data from distinct participant groups are inherently not interchangeable).

As in the mixed effects modelling of the behavioral data, *Word* and *Participant* served as crossed random effects (intercepts), and we tested for random slopes of the predictors of interest *Task* and *Speaker match condition* (*Task*, *Speaker match condition* and their interaction by-word, and *Speaker match condition* by-participant), using the same inclusion criteria as in the behavioral analyses. For the best model, we removed data points with standardized residuals exceeding 2.5 standard deviation units and refitted the model.

We ran the analyses on both the full set of trials (as recommended by VanRullen, 2011) as well as on the subset of trials that received correct answers. In all of our analyses, these two sets of data yielded similar results. We therefore only report analyses ran on the trials which received correct responses (in line with the log RT analyses).

Results and discussion

To make sure we only considered data from participants who were engaged in the tasks in a serious manner, we excluded one participant who showed an excessive error rate (i.e., 50% of errors) in the semantic classification task. Five participants were excluded due to technical issues, and a further sixteen participants were discarded due to excessive artefacts in their EEG data (affecting over 30% of trials). This left 45 participants in the final dataset, 22 in the old-new judgment task group (6 male; 18 - 30 years old; mean: 22 years old), and 23 in the semantic classification task group (3 male; 19 - 30 years old; mean: 22 years old).

Behavioral results

Response accuracy

Participants' behavior in the familiarization phases is summarized in Table 3.3. As the loudness task is highly subjective in nature, assessing correctness of the responses is not possible. We therefore report the average proportion of response agreement for participants receiving the same experimental lists.

On the targets, presented in the test phases, participants made on average 13% of errors. In the old-new judgment task, the mean error rate was 24%, while in the semantic classification task, participants' performance was near ceiling (2% of errors on average). Figure 3.1 displays the behavioral data (response accuracy and RTs) for both tasks as a function of *Speaker match condition*.

Table 3.3: Participants' behaviour in the familiarization phases of the experiment, displayed per test task. In the familiarization phases, participants carried out a loudness judgment task. *SE*s of the means for the RTs are given in brackets.

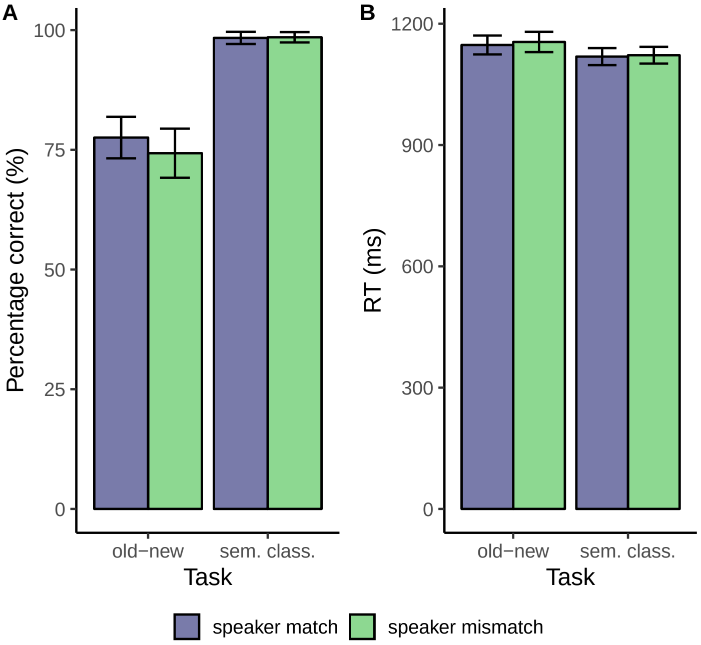
	RTs (ms)	**Proportion agreement**
Old-new judgment	1500 (14)	0.59
Semantic classification	1503 (13)	0.64

Participants' errors (pooled over tasks) could be predicted by the control predictor *Lag* ($\hat{\beta}$ = -0.02, t = -4.6, p < .001), showing that more errors were made on targets further away from their primes. In addition, we found an effect of *Word frequency* ($\hat{\beta}$ = -0.34, t = -3.3, p < .001), indicating that more errors were made on words with higher frequencies of occurrence. This frequency effect was likely mostly driven by the old-new judgment task, as words with higher frequencies of occurrence are known to appear inherently more familiar to participants, making these words susceptible to incorrect 'old' judgments in old-new judgment tasks (e.g., Glanzer & Adams, 1985). Moreover, we found that error rates were significantly higher for the old-new judgment task than for the semantic classification task ($\hat{\beta}_{\text{old-new task}}$ = -3.28, t = -10.2, p < .001, relative to the semantic classification task that is on the intercept). More importantly, *Speaker match condition* was not significant, neither as simple effect, nor in an interaction with *Task*.

RTs

RTs to correctly classified targets (measured from word onset, log-transformed for statistical analysis, 5% of RT outliers that were > 2 SD away from the grand mean removed) ranged from 650 to 2004 ms, and were 1133 ms on average (*SD* = 280 ms). In the old-new judgment task, RTs ranged from 660 to 2057 ms, and were 1160 ms on average (*SD* = 289 ms), while in the semantic classification task, RTs ranged from 650 to 1973 ms, and were 1117 ms on average (*SD* = 276 ms).

Figure 3.1: Mean percentage of errors (panel A) and mean RTs (panel B) to targets in the two tasks, split per speaker match condition. Error bars represent 95% confidence intervals. ‘Old-new’ denotes the old-new judgment task, ‘sem. class.’ the semantic classification task.



The log RTs (pooled over task) showed effects of the control predictors *Word duration* ($\hat{\beta}$ = 0.31, t = 10.8), *log RT preceding* ($\hat{\beta}$ = 0.02, t = 2.3), *log RT prime* ($\hat{\beta}$ = 0.02, t = 2.6), *Lag* ($\hat{\beta}$ = 0.001, t = 2.7), and *Part* ($\hat{\beta}_{\text{part 2}}$ = -0.06, t = -8.4; relative to part 1 on the intercept). These effects show that responses were quicker for targets with shorter durations, for targets whose preceding stimuli or primes received quick responses, for targets occurring more closely to their primes, and for targets occurring in the second part of the experiment. More importantly, the interaction of interest between *Task* and *Speaker match condition* was not statistically significant. The simple effects of these two predictors were not statistically significant either.

ERP results

100-300 ms: posterior sites

In the 100-300 ms time window on posterior electrode sites, we investigated exemplar effects on the N100/P200 complex, which indexes early acoustic processing. Exemplar effects in previous studies on this component consisted of more negative amplitudes for the match than for the mismatch condition. The average amplitudes in this time window were negative for all conditions and tasks, but there were relative amplitude differences between conditions and tasks (see Figure 3.2).

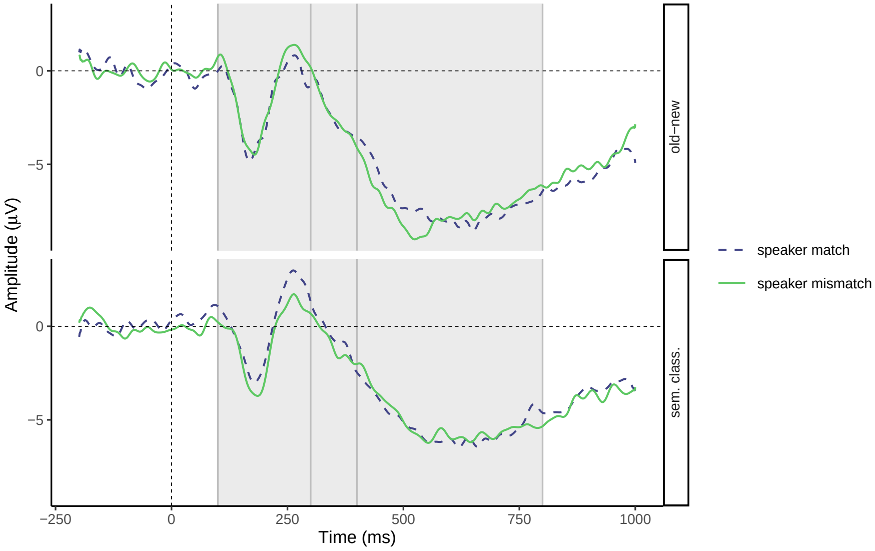
We found a statistically significant interaction between *Task* and *Speaker match condition*. This interaction appeared in our linear mixed effects models (see Table 3.4) as well as in an ANOVA: *Task* ($F(1, 43) = 0.91$, $p = .34$); *Speaker match condition* ($F(2, 86) = 0.53$, Greenhouse-Geisser adjusted $p = .57$); *Task* x *Speaker match condition* ($F(2, 86) = 4.78$, Greenhouse-Geisser adjusted $p < .05$). The *Speaker match condition* effects for both tasks in this and the other two time windows are illustrated in Figure 3.2 at central electrode C4 (an electrode representative for each of the ERP components tested in our study).

Table 3.4: Linear mixed effects model of amplitudes elicited by correct targets in the three time windows. The intercept represents speaker mismatch and the old-new judgment task. Estimated standard deviation is denoted by *SD*, the old-new judgment task by 'old-new', and the semantic classification task by 'sem. class'.

	100-300		**300-400**		**400-800**	
Fixed effects	$\hat{\beta}$	*t*	$\hat{\beta}$	*t*	$\hat{\beta}$	*t*
Intercept	-1.07	-3.6	-1.93	-5.4	-5.64	-17.4
Task (sem. class.)	0.35	0.9	0.60	1.3	-	-
Speaker match condition (match)	-0.36	-4.8	-0.17	-1.9	-	-
Task (sem. class.) * Speaker match condition (match)	0.56	5.8	0.28	2.4	-	-
Random effects		*SD*		*SD*		*SD*
Participant (intercept)		1.16		1.52		2.11
Word (intercept)		0.11		1.15		0.93
Task (old-new)		1.28		-		-
Task (sem. class.)		1.16		-		-

We ran linear mixed effects analyses on the data split according to task to further investigate the *Task* x *Speaker match condition* interaction. These models (without the simple and interaction fixed effects of *Task*, and without the by-word random effect of *Task*) show that for the old-new judgment task, amplitudes for the speaker match condition were significantly more negative compared to amplitudes in the speaker mismatch condition (speaker match condition: -1.40 μV on average; speaker mismatch condition: -1.15 μV on average; $\hat{\beta}_{\text{speaker match condition}}$ = -0.35, *t* = -4.7 relative to speaker mismatch condition on the intercept; also see Figure 3.2). For the semantic classification task, we observed the opposite pattern: amplitudes in the speaker match condition were significantly less negative compared to amplitudes in the speaker mismatch condition (speaker match condition: -0.49 μV on average; speaker mismatch condition: -0.74 μV on average; $\hat{\beta}_{\text{speaker match condition}}$ = 0.21, *t* = 3.5; also see Figure 3.2).

Figure 3.2: ERPs at electrode C4 for the old-new judgment (top) and semantic classification tasks (bottom). The three analyzed time windows (100-300 ms, 300-400 ms, and 400-800 ms) are highlighted in gray. 'Semantic class.' denotes the semantic classification task; negativity is plotted downwards.



300-400 ms: posterior sites

In the 300-400 ms time window on posterior electrode sites, we tested for exemplar effects on the P350 component. Previous studies established a reduction of the P350 (i.e., less positive amplitudes) for the match compared to the mismatch condition.

As in the 100-300 ms window, we obtained negative average amplitudes for all conditions and all tasks in this window (see Figure 3.2). Relative to these overall negative amplitudes, a reduced P350 (as reported in the literature for the match condition) consists of more negative amplitudes, and an enhanced P350 of less negative amplitudes.

The ANOVA analyses showed no interaction between *Task* and *Speaker match condition*. The linear mixed effect models, in contrast, detected a statistically significant interaction between *Task* and *Speaker match condition* (see Table 3.4). This interaction indicates that the old-new judgment task showed exemplar effects as in the literature (i.e., more negative amplitudes in the speaker match than the speaker mismatch condition) more clearly than the semantic classification task did.

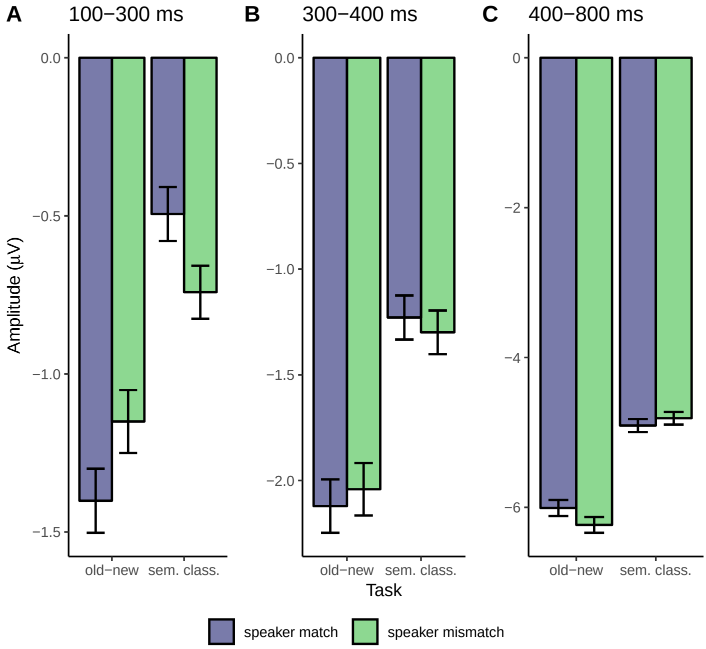
We applied linear mixed effects analyses to the data split according to task (without the simple and interaction fixed effects of *Task*) to inspect the *Task* x *Speaker match condition* in more detail. On either task analyzed separately, we found that differences between speaker match and speaker mismatch were not statistically significant (old-new judgment task: $\hat{\beta}_{\text{speaker match condition}}$ = -0.18, t = -1.9; semantic classification task: $\hat{\beta}_{\text{speaker match condition}}$ = 0.12, t = 1.6), probably due to a lack of statistical power.

400-800 ms: whole brain

In the 400-800 analysis window, we tested for exemplar effects on the N400 component over all electrode sites. Exemplar effects on this component are expected to surface as reductions of the N400 (i.e., less negative amplitudes) for the match condition compared to the mismatch condition.

Amplitudes in this time window were negative in all conditions (as was the case in the previous two analysis windows), but differed in relative amplitude according to task and condition. We observed that for the old-new judgment task, amplitudes were less negative in the speaker match than in the speaker mismatch condition (i.e., the expected direction for exemplar effects). For the semantic classification task, instead, amplitudes were slightly more negative in the speaker match condition than in the speaker mismatch condition (see Figure 3.3). However, this difference in the effect of *Speaker match condition* between the two tasks was not supported by a statistically significant *Task* x *Speaker match condition* interaction in our analyses (i.e., neither in the ANOVAs nor in the LMERs; the latter reported in Table 3.4).

Figure 3.3: Mean amplitudes of correctly classified targets in the three time windows under investigation (100-300 ms, 300-400 ms, and 400-800 ms), split according to speaker match condition and task. Note that the scaling of the y-axis differs per time window. Error bars represent 95% confidence intervals. 'Sem. class.' denotes the semantic classification task, 'old-new' the old-new judgment task.



General discussion

Which experimental tasks favor exemplar effects is informative about the nature and role of exemplars in speech comprehension. A number of previous studies have investigated this issue, but taken as a whole, their results remain inconclusive. In the present study, we examined whether exemplar effects arise more reliably in tasks that are explicitly memory-based versus tasks that are not. As a secondary research question, we investigated whether task-dependent exemplars effects are more adequately measured with EEG than with behavioral measures.

We carried out a long-term auditory repetition priming experiment, in which primes and targets were spoken in the same or a different speaker voice. Participants in our experiment engaged in either an old-new judgment task or in a semantic classification task on the targets; only the old-new judgment task requires participants' explicit reliance on episodic memory. We tested whether participants' processing in each task differed depending on whether targets were presented in the same or the other voice as their primes by collecting participants' RTs, response accuracies, and their EEG signals.

Preceding ERP literature established exemplar effects in ERPs on the P100/N200, the P350, and N400 components, respectively associated with acoustic processing, lexical activation, and lexical access. We tested for exemplar effects in time windows and scalp topographies corresponding to those components (i.e., 100-300 ms on posterior electrodes, 300-400 ms on posterior electrodes, and 400-800 ms on all electrodes, respectively).

In the ERPs, we found the clearest differences between the speaker match and mismatch conditions in the time window ranging from 100 to 300 ms, the window in which we tested for exemplar effects on the N100/P200 components. For the old-new judgment task, we established a difference between speaker match and speaker mismatch in line with the literature: like Schild et al. (2014b), we found more negative amplitudes for the match compared to the mismatch condition over posterior electrode sites. In contrast to the speaker voice variation tested in our study, the exemplar effects established by Schild and colleagues were for variation in stress (i.e., stressed or unstressed spoken prime syllables preceded spoken target words with initially stressed syllables). As in our study, a study by Campeanu et al. (2014) tested for exemplar effects of speaker voice. Like us and Schild and colleagues, Campeanu et al. established exemplar effects early in processing (i.e., from 200 ms onwards). However, as the scalp distribution of the exemplar effects observed by Campeanu and colleagues (i.e., over left anterior electrodes) clearly differed from the topographical distribution of the exemplar effects found by us and Schild and colleagues (i.e., over bilateral posterior electrodes), it appears that the exemplar effects reported by Campeanu and colleagues were not modulations of the early acoustic components (as the authors themselves note, too). Our study is therefore the first to observe a modulation of the N100/P200 complex as exemplar effects of speaker voice.

An earlier eye-tracking study indicated that exemplar effects of speaker voice may be most robust in measurements which precede participants' overt responses (Papesh et al., 2016). Exemplar effects were largest in the earliest measure of that study, which captured participants' saccade initation times, occurring around 700 ms after word onset. Our study, Campeanu et al. (2014), and Dufour et al. (2017) extend these findings by showing that in ERPs, exemplar effects of speaker voice may start a great deal earlier in processing than at saccade initiation times: from 100 ms after word onset in our study (i.e., in the 100-300 ms analysis window), from 200 ms in the study by Campeanu and colleagues, and from 450 ms in the study by Dufour and colleagues.

As for the old-new judgment task, we found a difference between speaker match and speaker mismatch for the semantic classification task. However, difference between speaker match and speaker mismatch for the semantic classification task went in the other direction as for the old-new judment task and as in the study of Schild et al. (2014b): the speaker match condition produced *less* negative amplitudes than the speaker mismatch condition. Despite its unforeseen direction, this significant difference between the speaker match and mismatch conditions in the semantic classification task suggests that, as in the old-new judgment task, participants' processing differed depending on speaker match condition in the semantic classification task. In itself, this finding may signify that exemplar effects arose. If so, both of our tasks elicited exemplar effects, but the way these exemplar effects were expressed in the ERPs was highly distinct depending on participants' task.

An alternative interpretation of the findings for the semantic classification task is that they are *opposite* effects from exemplar effects. This would mean that a facilitation in processing arose in the mismatch compared to the match condition. In behavior, opposite effects have been reported too, by us (reported in Chapter 2 of this dissertation), and by Morano et al. (in press), for example. In these studies, participants' responses were faster in the *mismatch* conditions. These mismatch effects were interpreted as not reflecting participants' use of exemplars, but rather their use of abstract lexical representations. As a result of specific characteristics of targets' primes (for example their relative intelligibility), targets' abstract representations in the mismatch condition were believed to receive higher levels of activation. For instance, in Morano et al. (in press), less intelligible primes, occurring in the mismatch condition, took participants longer to process. Consequently, the targets' abstract lexical representations received more activation which sped up their recognition.

Primes in the present study were spoken by a male (in the speaker match condition) or a female speaker (in the speaker mismatch condition), while targets were always uttered by the male speaker. Even though both the male and female talker in our study are normal (e.g., non-dialectal) and healthy, a rating study on tokens produced by these same talkers (reported in Chapter 2) showed that the female speaker is slightly less intelligible than the male speaker. Therefore, if more negative amplitudes in the 100-300 ms window index a form of processing facilitation for one experimental condition relative to another (as was likely the case in the speaker match condition for the old-new judgment task), it is possible that the more negative amplitudes we found for the speaker mismatch than for the speaker match condition in the semantic classification task reflect higher activation of targets' abstract representations due to the relatively low intelligibility of the female speaker's primes, occurring in the speaker mismatch condition. The difference between speaker match and speaker mismatch we observed for the semantic classification task would then not be exemplar effects, but rather stem from increased activation of targets' abstract representations.

The following EEG analysis window ranged from 300-400 ms, and tested for exemplar effects on the P350 component. Preceding literature by Friedrich, Kotz, Friederici, and Alter (2004) and Schild et al. (2014b) established reductions of the P350 (i.e., less positive amplitudes) in the match compared to the mismatch condition for word stress. Friedrich and colleagues interpreted these reductions of the P350 for stress-matching targets as reflecting facilitated lexical identification (while Schild et al., 2014b linked their prolonged effects to the N400). In this time window, we established overall negative amplitudes. We found an interaction between *Task* and *Speaker match condition* (in the LMER but not the ANOVA analyses), which indicated that the speaker match condition elicited more negative amplitudes than speaker mismatch (i.e., exemplar effects as expressed on overall negative amplitudes) especially for the old-new judgment task. It therefore appears from our study that a match in speaker voice, like a match in word stress, may result in enhanced facilitation of lexical identification depending on participants' task.

The final EEG time window we analyzed (400-800 ms, over all electrode sites) aimed to capture exemplar effects on the N400 component, a well-described ERP component which has been related to lexical access. In lexical priming paradigms, the N400 is reduced (i.e., less negative amplitudes arise) for targets that better match their primes (e.g., Kutas & Van Petten, 1988). In Dufour et al. (2017) and Schild et al. (2014b), the N400 was respectively sensitive to a match in speaker voice or in word stress (although the polarity of the effects was not compatible with the expected direction from the N400 literature in all experiments). In our data, we found no statistically significant *Task* x *Speaker match condition* interactions. Numerically, however, we observed differences between our tasks. The old-new judgment task showed the expected pattern: the speaker match condition produced a reduced N400 compared to the speaker mismatch condition (i.e., less negative amplitudes). In the semantic classification task, instead, the speaker match condition yielded a slightly enhanced N400 (i.e., more negative amplitudes) compared to the speaker mismatch condition.

The fact that the *Task* x *Speaker match condition* interaction did not reach statistical significance in the 400-800 ms analysis window may stem from our choice of electrode sites. The choice of electrode sites for analysis affects the strength of exemplar effects on the N400; in the few studies that showed exemplar effects on the N400, the effects appeared to be somewhat unstable depending on which electrode sites were analyzed precisely. We deliberately chose our analysis electrode sites based on the wide centro-parietal topographical distribution known from the substantial general literature on the N400 (e.g., Kutas & Federmeier, 2011), and carefully avoided anti-conservative follow-up searches for partial results by specific topographical subsetting. Another possible reason for the null results in the 400-800 window may be the position of the left boundary of the window. Closer inspection of the ERPs of representative central electrodes (of which one example is displayed in Figure 3.2) suggests that differences between the speaker match and mismatch conditions on the N400 for the old-new judgment task began earlier than 400 ms. It is therefore possible that we missed part of the effects because they fell outside of the analysis window.

Our first research question investigated whether exemplar effects differ between tasks which do and do not require participants' explicit use of episodic memory. Our results are in line with previous literature (e.g., Luce & Lyons, 1998), and show that exemplar effects mostly arise when participants engage in a task that requires their use of episodic memory. Importantly, this relationship between participants' use of episodic memory and the occurrence of exemplar effects may be indicative for a crucial role of episodic memory in the mental representation of exemplars.

Our second research question focused on experimental methods, and investigated whether exemplar effects are more adequately captured with EEG than with behavioral measures. Importantly, we found no exemplar effects in the behavioral data for either task, while we observed clear differences between the match and the mismatch conditions in the EEG data. Our study therefore demonstrates that ERPs provide a more sensitive assessment of exemplar effects than behavioral methods do. The exemplar effects we found in the ERPs were likely too small in size to be reliably captured by the behavioral responses, which reflect an accumulation of processes (i.e., these exemplar effects could have been washed away by other effects also captured in the behavioral responses). The difference in exemplar effects we established between the ERPs and behavioral measures also suggests that in some of the experiments in the behavioral literature that produced null results, exemplar effects may have been present, but were too weak to be detected by the behavioral measures used.

ERPs also have the advantage that they can provide more information about the point in time at which exemplars play a role. The ERPs in the old-new judgment task showed that exemplar effects especially arise early in processing. Which exact functional aspects of processing are affected by exemplars is less clear, however, as the N100 and P200 components captured in the 100-300 ms window (in which the clearest exemplar effects arose) are associated with multiple other functional aspects than early acoustic processing. For instance, the N100 has also been linked to attention allocation (e.g., Hink, Van Voorhis, Hillyard, & Smith, 1977), working memory operation (e.g., Golob & Starr, 2004), and listeners' mental state (e.g., Näätänen & Piction, 1987), and the P200 to stimulus expectancy (e.g., Federmeier & Kutas, 2002) and rapid attention switching (e.g., Furutsuka, 1989), among other things. As a result of these various functions associated with ERP components, it is not always easy to functionally interpret a modulation of an ERP component for a certain experimental condition, especially in the absence of a precedent in the literature.

In summary, our ERP measurements but not our behavioral findings show reliable differences in exemplar effects between tasks. Exemplar effects clearly arose early in processing for an old-new judgment task, while reversed effects were established for a semantic classification task. These reversed effects are either exemplar effects expressed differently, or reflect participants' use of abstract representations. For both scenarios, we can conclude that exemplar effects play a different role when listeners rely on their episodic memories than when they do not: our findings suggest a relationship between the representation of exemplars and episodic memory.

Exemplar effects arise in a lexical decision task, but only under adverse listening conditions

Chapter 4

This chapter is a reformatted version of
Annika Nijveld, Louis ten Bosch and Mirjam Ernestus (2015). Exemplar effects arise in a lexical decision task, but only under adverse listening conditions. *Proceedings of the International Congress of Phonetic Sciences 2015*, Glasgow, Scotland.

Abstract

This paper studies the influence of adverse listening conditions on exemplar effects in priming experiments that do not instruct participants to use their episodic memories. We conducted two lexical decision experiments, in which a prime and a target represented the same word type and could be spoken by the same or a different speaker. In Experiment 1, participants listened to clear speech, and showed no exemplar effects: they recognized repetitions by the same speaker as quickly as different speaker repetitions. In Experiment 2, the stimuli contained noise, and exemplar effects did arise. Importantly, Experiment 1 elicited longer average RTs than Experiment 2, a result that contradicts the time-course hypothesis, according to which exemplars only play a role when processing is slow. Instead, our findings support the hypothesis that exemplar effects arise under adverse listening conditions, when participants are stimulated to use their episodic memories in addition to their mental lexicons.

Introduction

Hybrid models of speech comprehension assume two types of lexical representations for the pronunciation of words: abstract representations and exemplars (e.g., Goldinger, 2007; Hawkins, 2003). Abstract representations are strings of sound symbols that contain no details about the exact pronunciations of words. Exemplars, in contrast, are highly detailed representations of each occurrence of a word, which together form a word cloud. Findings from auditory priming experiments offer support for the representation of words as clouds of exemplars (e.g., Bradlow et al., 1999; Craik & Kirsner, 1974; Palmeri, Goldinger, & Pisoni, 1993). In these studies, listeners recognized the second occurrence of a word more quickly or more accurately when surface details (e.g., speaker voice) of the first (prime) and second (target) occurrence of a word matched compared to when they did not match. However, several studies did not replicate these exemplar effects (e.g., Goldinger, 1996; McLennan & Luce, 2005). Because the precise role of exemplars has important consequences for models of speech comprehension, we further investigated under which circumstances exemplars play a role in speech comprehension.

We tested the hypothesis that exemplar effects arise when participants do not only use their mental lexicons to perform a task but also their episodic memories. They may do so because the task in an experiment instructs them so (e.g., old-new judgment) or because relying on episodic memory offers them a processing advantage, for example, under adverse listening conditions.

Exemplar effects indeed arose in previous priming studies where listening conditions were suboptimal: in Mattys and Liss (2008), participants listened to healthy, mildly dysarthric or severely dysarthic speakers, and words were either repeated by the same or a different speaker. Participants were faster to recognize same compared to different voice repetitions, but only when listening to mildly or severely dysarthic speech, and mostly so in the latter case. In Saldaña et al. (1996), words were presented in clear speech, or with increasing levels of white noise (+5 to -5 SNR). The authors found larger exemplar effects on accuracy when the signal to noise ratio was lower.

In these studies, not only were listening conditions suboptimal, but the task (old-new judgment) also instructed participants to use their episodic memories. It is therefore unclear which aspect caused the exemplar effects (and to what extent).

In Hanique et al. (2013), participants listened to words produced in full or with the first-syllable schwa missing, by the same speaker. Participants were faster to recognize repeated words in the same than in a different pronunciation variant. The variants without schwa occur frequently in casual speech but seldom out of context as in Hanique et al. (2013). Their presence in the experiment may therefore have increased participants' processing load, who were thus stimulated to use their episodic memories. The effect in McLennan and González (2012), where participants listened to a talker with a foreign accent, likely also arose for this reason, but was confounded with speaker.

We report two experiments to test our hypothesis that if participants are not instructed to use their episodic memories, exemplar effects especially arise under adverse listening conditions. Experiment 1 is based on the lexical decision experiment by Hanique et al. (2013) described above, but we used clear speech produced by two different speakers instead of two pronunciation variants produced by one speaker. If the exemplar effects reported by Hanique et al. (2013) result from the adverse listening conditions, as we hypothesize, our version of the experiment should not show exemplar effects. In Experiment 2, we presented the same stimuli but added speech-shaped noise (+3 dB SNR) to them, which made them harder to understand. We predict that this experiment will show exemplar effects again.

Experiment 1

Method

Participants

We tested 26 participants, aged between 18 and 25 years (mean: 20 years). Five were male and two were left-handed. None of the participants in this or the other experiment presented in this paper reported any hearing impairment or participated in both experiments. All were paid for their participation.

Materials

The lexical decision experiment was run in Dutch and contained the same words and pseudo words as Hanique et al. (2013); all were trisyllabic prefixed infinitives (starting with *be-* or *ver-*, e.g. real word *vertellen* ‘to tell’ and pseudo word *bekrempen*, see also the Appendix). All pseudo words were phonotactically legal in Dutch. We repeated 48 real infinitives, which had a mean frequency of occurrence of 3362 per million (range: 456 - 8296 per million; based on Baayen et al., 1995).

We divided the experiment in two parts, where each part contained both tokens (prime and target) of half of the repeated words. Each part consisted of two blocks: a familiarization block, with 24 primes, 24 to-be-repeated pseudo word foils and 24 additional foils (12 pseudo words; 12 real words), and a target block, with 24 targets, 24 repeated pseudo word foils and 24 additional foils (12 pseudo words; 12 real words). One real word foil (*besmetten*, ‘to infect’) was accidentally repeated in Hanique et al. (2013); we replaced its second occurrence by *bestijgen* (‘to ascend’).

We recorded all items with a male and a female native speaker of Dutch in a sound-attenuating booth. Target words spoken by the male speaker had an average duration of 688 ms (range: 580 - 843 ms; *SD:* 56 ms), while the targets spoken by the female speaker were 626 ms on average (range: 469 - 760 ms; *SD*: 67 ms).

For the presentation of the trials to participants, we created four lists, in which words occurred in the same pseudo randomized order as in the four master lists of Hanique et al. (2013). In the lists, each block started with a foil, primes and targets were always followed by a foil, at most eight real or pseudo words occurred in succession, and primes and targets were separated by maximally 100 trials (average: 67, range: 19 - 100). In each list, an equal number of prime-target pairs were assigned to one of the four possible combinations of speaker voice. Per master list, we created three lists with the same word order but in which the prime-target pairs represented one of the three other possible combinations of speaker voice. In each set of four lists, every prime-target pair represented each voice combination exactly once. In each block of each of the lists, half of the targets and approximately half of the foils were spoken by the male speaker, and the other half were spoken by the female speaker.

Procedure

Participants were tested individually in a sound-attenuating booth. They listened to the stimuli via closed headphones and performed a lexical decision task by means of button presses on a button box (yes-responses with the dominant hand). Per trial, one stimulus was presented and the next trial started one second after a button press or after 3.5 seconds after trial onset. Between the two parts of the experiment, participants took a short break, and one session lasted approximately 15 minutes.

Analyses

We analyzed log-transformed response times (RTs) to the target words by means of mixed effects regression models, and accuracy scores of words and pseudo words by means of generalized mixed effects regression models. We used word, participant and speaker (of the target word) as crossed random effects in both analyses. We restricted the RT analysis to trials that received correct responses and whose primes also received correct responses, and we removed trials with response times that differed more than two standard deviations from the grand mean. We removed participants and words who were outliers relative to the mean in each experiment in terms of error rate or in terms of error and missed response rate from the analyses.

For both the RT and accuracy analyses, we tested random slopes for all fixed effects. We only included effects and interactions if they were significant (as this type of data with many observations approaches a normal distribution, *t*-values above 1.96 or below -1.96 imply *p*-values < .05) and if they significantly improved the statistical model (tested with the $\mathrm{anova}()$ function from the R Statistical Software RCoreTeam, 2018). We orthogonalized correlating control variables before they were included in the model (*RT prime* was residualized over *RT preceding trial* in the RT analysis). For the RT analysis, we removed data points with standardized residuals exceeding 2.5 standard deviation units for the best model and refitted the model. Our main predictor was *speaker match*, which reflects whether or not a prime and target were pronounced by the same speaker. We also explored the influence of the control predictors log-transformed response times to the prime *(RT prime)* and to the preceding trial *(RT preceding trial)*, log-transformed *target word duration* and *affix* (whether target words carried the affix *be-* or *ver-*).

Results

Participants, on average, made errors on 9% of target words and pseudo words (which includes missed responses, 4% of responses were incorrect). Analysis of accuracy scores on the target words did not reveal any effect of *speaker match*. Participants' behavior per speaker match condition in both experiments is summarized in Table 4.1.

Table 4.1: Participants' behavior to target words split by *Speaker match* in Experiments 1 and 2. For the mean RTs to correctly classified targets, 95% confidence intervals are given in brackets. Probability correct on targets and fillers (which includes missed responses) is given with the lower and upper boundaries of Wilson's 95% confidence intervals. 'Match' and 'mismatch' refer to the speaker match- and mismatch conditions.

	RT (ms)		**Probability correct**	
	Match	Mismatch	Match	Mismatch
Experiment 1	974 (14)	983 (14)	0.92 (0.89-0.94)	0.94 (0.92-0.95)
Experiment 2	918 (14)	933 (14)	0.90 (0.87-0.92)	0.89 (0.86-0.91)

For the response time (RT) analysis, we excluded one participant (13% response errors) and one word (*bekransen* 'to garland': 48% response errors). Table 4.2 shows the statistical model based on the remaining 986 trials[1]. Response times, measured from word onset, were on average 986 ms. The effects of our control predictors reveal that participants responded more quickly to targets when they were quicker on the preceding trial or on the prime, when the word started with *be-* (mean: 952 ms) rather than *ver-* (1006 ms) and when the word was shorter. Crucially, we did not observe a significant effect of *speaker match*, which indicates that participants were not quicker to respond to targets pronounced by the same speaker as the prime than to targets pronounced by the other speaker.

Experiment 2

Method

Participants

The participants were 26 native speakers of Dutch, aged between 18 and 29 years (mean: 22 years). Five were left-handed and seven were male.

Materials, procedure and analyses

We used the same materials as in Experiment 1, but superimposed speech-shaped noise at +3 dB SNR to the recordings (energetic masking). An informal pre-test confirmed that this noise level made the words harder to identify while it was still possible to perform the task. We first modified the loudness of the speech to reach the desired SNR level, and subsequently rescaled the speech + noise to the original loudness level of the speech (70 dB). The noise started and ended with a 30 ms ramp.

The procedure was identical to the one of Experiment 1. For the statistical analysis, we used the same method and predictors as in Experiment 1.

Results

On average, participants made 12% errors on target words and pseudo words (which includes missed responses, 9% of actual responses was incorrect). The errors on target words showed no effect of *speaker match*. For the RT analysis, we again

[1] The results of these models are highly similar to the ones reported in the published paper, but based on slightly improved statistical modelling.

excluded the word (*bekransen* ‘to garland’: 70% response errors) as well as two participants (with respectively 23% response errors, and 34% of response errors and misses). The statistical model based on the remaining 891 trials is summarized in Table 4.2. Response times to target words were on average 925 ms. The control predictors that played a role in Experiment 1 were again significant and showed similar effects, except for the predictor *affix* that was no longer significant. Importantly, we found an interaction between *speaker match* and *RT prime*, which indicated that participants were quicker to respond to target words repeated by the same speaker than to words repeated by a different speaker, but only for those target words whose primes received quick responses.

Table 4.2: Statistical model for the response times to targets in Experiments 1 and 2. Estimated standard deviation is denoted by *SD*.

	Experiment 1		**Experiment 2**	
Fixed effects	$\hat{\beta}$	t	$\hat{\beta}$	t
Intercept	2.97	9.3	3.09	8.3
RT preceding trial	0.22	9.9	0.15	6.7
RT prime	0.12	5.4	0.14	4.0
Affix *ver-*	0.02	2.5	-	n.s.
Target duration	0.37	8.4	0.41	7.9
Speaker mismatch	-	n.s.	0.01	1.6
Speaker mismatch * RT prime	-	n.s.	-0.09	-2.1
Random effects		SD		SD
Word	intercept	0.02	intercept	0.04
Participant	intercept	0.07	intercept	0.06
Speaker	intercept	0.02	intercept	0.02
Residual		0.11		0.11

The difference between Experiments 1 and 2 in the effect of *speaker match* modulated by *RT prime* is supported by an analysis of the combined datasets, which shows a significant interaction between *speaker match*, *RT prime* and *experiment* ($\hat{\beta}$ = -0.12, t = -2.0). This result confirms that *speaker match*, modulated by *RT prime*, only played a role in Experiment 2.

General discussion

In this paper, we investigated the hypothesis that when participants are not instructed to use their episodic memories, exemplar effects arise especially under adverse listening conditions. To this end, we reported two experiments in which participants performed lexical decision, a task that does not require participants' use of episodic memory.

In Experiment 1, participants listened to clear speech produced by two different speakers. As hypothesized, we found no exemplar effects: participants were equally fast to recognize repeated words spoken by the same speaker and by the other speaker. This results contrasts with the lexical decision experiment reported by Hanique et al. (2013), which used the same words. The crucial difference between the two experiments is that Hanique et al. (2013) presented half of the stimuli with the first-syllable schwa missing. These casual pronunciation variants seldom occur in isolation, as they were presented in Hanique et al. (2013), and were therefore presumably hard to process. This result confirms our hypothesis that exemplar effects may arise under adverse listening conditions.

In Experiment 2, we re-introduced an adverse listening condition by adding noise to the stimuli of Experiment 1. In contrast to Experiment 1, we did find exemplar effects here: for targets whose primes received quick responses, participants were quicker to respond to words repeated by the same speaker than by the other speaker. Experiment 2 only differed from Experiment 1 in the noise that we added; this result therefore suggests that the exemplar effects in Experiment 2 arose because of the adverse listening condition that was created by the noise. This finding provides further support our hypothesis that adverse listening conditions stimulate participants to use their episodic memories, which evokes exemplar effects.

The exemplar effect we found in Experiment 2 was modulated by *RT prime*, which indicates that exemplars only played a role for those targets whose primes elicited quick responses. RTs for the primes are likely a measure of how easily participants could process the primes (Cutler & Robinson, 1992). A relatively long RT on a prime may reflect a case in which the participant - even though (s)he gave a correct response - found it hard to identify the prime through the noise. In these cases, participants may have missed details that are needed to build a full episode. Due to the prime's impoverished representation, participants may have been unable to match the target to the prime when they heard the target, and could not benefit from using episodic memory in these cases.

The RTs in Experiment 2 were shorter (925 ms) than in Experiment 1 (986 ms), a result that may be surprising given the enhanced listening difficulty in Experiment 2. Possibly, the more challenging stimuli in Experiment 2 made participants in this experiment more motivated, a factor that can lead to faster responses (Capa, Audiffren, & Ragot, 2008).

The time-course hypothesis (McLennan et al., 2003; McLennan & Luce, 2005) predicts that exemplar effects only arise when responses are relatively slow, because it takes more time to activate exemplars than abstract representations. Our findings do not support this hypothesis, as we observed exactly the opposite pattern of results: we only found exemplar effects in the experiment with shorter RTs.

The combined results of our study and Hanique et al. (2013) show that different kinds of adverse listening conditions can lead to exemplar effects in lexical decision experiments: both an experiment that (in part) included casually produced stimuli and an experiment with noise superimposed onto all stimuli showed exemplar effects. Interestingly, the effect in Hanique et al. (2013) (19 ms) was larger than the one in Experiment 2 (15 ms; collapsed over targets with fast and slow primes). Possibly, our experiment showed a smaller exemplar effect because schwa deletion in isolation represents a more severe adverse listening condition than moderate noise does.

Our results have implications for models of spoken word recognition. We found further support for the storage of word pronunciations as clouds of exemplars in Experiment 2, while the absence of exemplars effects in Experiment 1 favors storage as abstract representations. Hybrid models combine both types of representations, but cannot straightforwardly explain why we observed exemplar effects under adverse listening conditions only. Our results are therefore best explained by models that assume that during speech comprehension, the mental lexicon, containing abstract representations, co-operates with domain-general episodic memory, in which clouds of exemplars are represented. Depending on the situation, one or the other plays a more important role.

In conclusion, we investigated the influence of adverse listening conditions on the occurrence of exemplar effects, and only found exemplar effects in the experiment in which the stimuli contained noise. When no noise was used, no exemplar effects arose. This suggests that adverse listening conditions stimulate participants to use their episodic memories (also in the absence of a task that instructs them to do so), which enhances the probability of observing exemplar effects.

The use of exemplars differs between native and non-native listening

Chapter 5

This chapter is a reformatted version of
Annika Nijveld, Louis ten Bosch and Mirjam Ernestus (submitted). The use of exemplars differs between native and non-native listening.

Abstract

This study investigates the role of exemplars in native and non-native listening in two English lexical decision experiments. Participants were native English listeners, Dutch non-native listeners, and Spanish non-native listeners. In Experiment 1, primes and targets were spoken in the same or a different voice. The native listeners showed exemplar effects, while the non-native listeners did not. In Experiment 2, primes and targets were presented with the same or a different degree of vowel reduction. The Dutch, but not the Spanish listeners were familiar with this reduction from their L1 phonology. Exemplar effects only arose for the non-native listeners in this experiment, and exemplar effects were larger for the Spanish than the Dutch listeners. We conclude that the use of exemplars is modulated by factors such as listeners' availability of processing resources, listeners' familiarity with the variation type that forms the basis for the match/mismatch condition in an experiment from their L1 phonology, and the salience of the variation type. The use of exemplars differs between native and non-native listening, which suggests that there are qualitative differences between native and non-native speech comprehension processes.

Introduction

Hybrid models of speech comprehension distinguish two types of mental representations for the pronunciation of words: abstract representations and exemplars. Abstract representations consist of symbolic units (e.g., phonemes), and do not contain details about each token of a word (such as information about the speaker's gender, age, and mood). Exemplars, in contrast, mentally represent all experienced occurrences of a word in full phonetic detail, which together form a cloud associated with that word. Many studies in the literature point to a role for exemplars in spoken word recognition in listeners' *native* language (e.g., Goh, 2005; McLennan & Luce, 2005; Palmeri et al., 1993). This study investigates whether the use of exemplars differs between native (L1) and non-native (L2) listeners. Importantly, if the use of exemplars differs between native and non-native listening, there are qualitative differences between the weight and relative importance of speech processing mechanisms involved in these two types of listening.

The representation of words as exemplars for L1 listeners is, despite some mixed results, supported by a range of auditory identity priming studies (e.g., Craik & Kirsner, 1974; Goldinger, 1996; Pufahl & Samuel, 2014). In these experiments, participants recognized repeated words more quickly and/or more accurately if the two tokens of the word ('prime' and 'target') shared perceptual characteristics such as the speaker's voice (the 'match' condition) than when they did not (the 'mismatch' condition; these effects are referred to as 'exemplar effects'). The assumption is that all perceptual details of the prime are retained in memory as an exemplar, which affects the subsequent processing and recognition of the target.

Most of these experiments used variation in speaker voice as the basis for the match/mismatch condition, and therefore concluded listeners store this type of variation in the form of exemplars. Other variation types have been tested as well. For instance, exemplar effects have also been reported for variation in speech rate, emotional tone of voice, fundamental frequency, and the realization of a given single segment of a word (Church & Schacter, 1994; Janse, 2008; Krestar & McLennan, 2013; Sumner & Samuel, 2005). Native listeners may thus use exemplars with acoustic-phonetic information representing a range of variation types.

Whether also non-native listeners use exemplars in the speech recognition process is almost completely unknown. The processing load in L2 differs substantially from that in L1: while listening is remarkably effortless in the listeners' native language, this is not true for listeners' non-native language(s). Their increased processing load might affect listeners' reliance on exemplars. If exemplars play a substantial role in non-native listening, theoretical models of L2 word recognition would need to be adapted (e.g., the BIA+ model of Dijkstra & Heuven, 2002). So far, the assumption of a hybrid mental lexicon (with both abstract representations and clouds of exemplars) has hardly been adopted in the field of L2 language processing: Nearly all studies on non-native language processing assume that words are only stored as abstract representations (and thus not as clouds of exemplars). Our first research question is, therefore, whether the occurrence of exemplar effects differs between native and non-native listeners.

Two studies compared the occurrence of exemplar effects in native and non-native listening. Winters, Lichtman, and Weber (2013) tested for exemplar effects with German native listeners and English non-native listeners (with some or no knowledge of German). These listeners were presented with German words, and they indicated whether words occurred previously in the experiment or not (an old-new judgment task). Repeated (to be judged as 'old') words were presented in the same or in a different voice as during the words' first presentations (to be judged as 'new'). Both the native and the non-native listeners were more accurate to classify repeated words as 'old' when these were presented in the same voice as earlier. There was no difference between the listener groups; exemplar effects were thus of similar size between the groups. This study shows that L2 listeners may encode and use voice-specific information in speech recognition, just like native listeners.

In Trofimovich (2005), English listeners with self-reported 'low-intermediate proficiency' in Spanish were tested on familiar words in English (i.e., L1 stimuli) and Spanish (i.e., L2 stimuli). Listeners had to repeat out loud words that were presented auditorily to them; these prompts were presented in either the same or a different voice as during a familiarization phase earlier in the experiment. For the Spanish words (i.e., L2 stimuli), listeners were quicker to start producing the words played in the same voice than in the other voice. Hence, exemplar effects arose in non-native processing. It is unclear whether there was a difference between L1 and L2 processing because the difference between the size of exemplar effects for L1 and L2 stimuli was not statistically supported by an interaction between Language and Voice. In addition, the different stimulus words used between language conditions confounded the comparison between native and non-native processing. This study thus indicates that also L2 listeners may store information about speaker voice, but does not reliably show differences in exemplar effects between L1 and L2 listening.

A study by Drozdova, Hout, and Scharenborg (2019) tested for exemplar effects of speaker voice in noise and clean speech among Dutch non-native listeners of English. In this study, participants engaged in a word identification task and in an old-new judgment task. Exemplar effects arose in the old-new judgment task only, and the effects were larger for non-native listeners with higher proficiency levels. The intermediating effect of noise was less clear, as exemplar effects in accuracy were larger in noise, while in the RTs, the effects were larger in clean. This study also shows that non-native listeners may show exemplar effects for variation of speaker voice.

If exemplar effects arise in non-native listening, the question is what the role of listeners' L1 phonology is in the occurrence or size of exemplar effects. Non-native listeners show a decreased sensitivity to acoustic-phonetic information that is not distinctive in their native language (Kuhl, 2004), and their L1 phonology may impose a 'phonological filter' (Trubetzkoy, 1939) on perception. As a result of such insensitivity to part of the speech signal, non-native listeners may not show exemplar effects for variation unknown from listeners' L1 phonology. If so, the occurrence of exemplar effects may differ between non-native listeners with different L1 backgrounds. However, given the observation that non-native listeners proved sensitive to L1-irrelevant contrasts in certain tasks (Werker & Logan, 1985), it is also possible that the formation of exemplars is not bound by L1 phonology-imposed perceptual filters. Our second research question addresses this issue, and investigates how listeners' L1 phonology influences the occurrence or size of exemplar effects.

A study by Morano et al. (in press) is informative about the role of listeners' L1 phonology on the occurrence of exemplar effects in L2 listening. In that study, Dutch L2 listeners with intermediate proficiency levels engaged in a lexical decision task in French with pronunciation variants that matched or mismatched in the voicing of the vowel in words' initial syllables. Importantly, such voicing variation is not common in Dutch. Morano and colleagues conducted three versions of their experiment. In versions AA and BB, primes and targets in the match condition were identical tokens (where AA and BB represent two distinct sets of identical tokens). In version AB, primes and targets were always different tokens (i.e., both in the match and the mismatch conditions). Exemplar effects arose in version AA only. The occurrence of exemplar effects was thus highly restricted, and seemed to depend on specific stimulus characteristics. Nevertheless of relevance to the present investigation, exemplar effects arose for L1-*irrelevant* variation, which suggests that listeners' L1 phonology need not restrict the variation types for which exemplar effects arise.

In sum, whether there are differences in exemplar effects between native and non-native listening is an important but unresolved question. In addition, research is necessary to establish the role of listeners' L1 phonology for the occurrence or size of exemplar effects. To compare the occurrence of exemplar effects between native and non-native listening, we tested native and non-native listeners (English native listeners, and Dutch and Spanish non-native listeners of English) on a lexical decision task in English. We examined if the occurrence of exemplar effects is affected by non-native listeners' L1 phonology by testing non-native listeners from two backgrounds on a variation type that was familiar from the L1 phonology of one, but not the other group of non-native listeners.

In Experiment 1, as in the majority of the literature, we used indexical variation of speaker voice as the basis for the match/mismatch condition: repeated words were presented in the same or a different voice. This variation is L1-unspecific. All listeners have a great deal of experience dealing with this variation type from their L1. The experimental words were familiar British English words. This experiment investigates whether exemplar effects differ between native and non-native listeners for a common variation type.

In Experiment 2, we used variation stemming from speech reduction. Speech reduction is highly common in casual speech, and refers to the phenomenon whereby words' sounds are shortened and/or more weakly uttered, or absent compared to the sounds in words' citation forms (see Ernestus & Warner, 2011, for an introduction). The way speakers reduce words is partly language-specific. Native listeners comprehend reduced pronunciation variants (occurring in running speech) with remarkable ease. Reduced tokens in our experiment had shorter overall durations compared to unreduced tokens. More importantly, they had highly shortened vowels in their initial, unstressed syllables (e.g., *balloon* with a very short schwa). This English vowel reduction pattern is well-documented in the literature (e.g., Dalby, 1986; Shockey, 2003), and is highly frequent.

Unlike native listeners, non-native listeners have difficulties understanding speech reductions (e.g., Brand & Ernestus, 2018). Non-native listeners typically lack exposure to reduced pronunciation variants, which results in a lack of familiarity with these forms. However, non-native listeners can be familiar with reduction patterns through their L1, and an overlap between L1 and L2 reduction patterns may aid the perception of reduced pronunciation variants in L2 (e.g., Mitterer & Tuinman, 2012). Vowel reduction, as we studied in Experiment 2, is also common in Dutch, although unlike in English, vowel reduction in Dutch seems to result from categorical as well as gradient processes rather than just from gradient processes as in English (e.g., Bürki & Gaskell, 2012; Ernestus, 2000). Spanish, in contrast, hardly reduces its vowels (instead, consonants are reduced in that language; Torreira & Ernestus, 2011). In addition, the schwa vowel (as in most of our experimental words) is not part of Spanish phonology. Importantly, the Dutch but not the Spanish listeners are therefore familiar with the schwa vowel itself as well as with its reduction from their native language phonology.

Experiment 2 is thus informative about the types of variation for which non-native listeners show exemplar effects. If L1 phonology has a negligible influence on the perception and the encoding of the stimulus variation into memory, the occurrence of exemplar effects should not differ between the Dutch and the Spanish non-native listeners. Instead, if L1 phonology does have an effect, exemplar effects for the stimulus variation as applied in this experiment is expected to differ between the Dutch and the Spanish non-native listeners, since only the Dutch listeners are familiar with the reduction pattern from their native language phonology.

Experiment 1

Method

Participants

One-hundred and thirteen participants took part in the experiment (which excludes two participants whose data we could not use due to technical issues). Of these participants, 40 were native speakers of English (mean age: 21 years; 6 left-handed; 11 male), 40 were native speakers of Dutch (mean age: 21 years; 6 left-handed; 6 male), and 33 were native speakers of Spanish (mean age: 22 years; 5 left-handed; 21 male). All participants were highly educated, reported no hearing disorders, gave their informed written consent, and were paid for their participation.

We assessed our non-native listeners' English proficiencies with the LexTALE task (Lemhöfer & Broersma, 2012), on which the Dutch listeners obtained an average score of 74%, while the Spanish listeners were at 67% on average. Both of these averages fall within CEFR level B2 ('upper intermediate' proficiency – 60-80%). Self-rated English proficiency (on a 1 - 6 scale) was at 4.8 on average for the Dutch listeners (*SD* = 1.0) and at 4.3 for the Spanish listeners (*SD* = 0.9).

Materials

The experiment contained 43 bi- and trisyllabic real English nouns with stress on the second syllable, and an equal number of counterpart pseudo words (listed in the Appendix). We derived the pseudo words from the real words by keeping the initial syllable, and altering up to three phonemes in the following syllables through substitution or deletion (e.g., we derived pseudo word *ballee* from real word *balloon*). This procedure resulted in pseudo words with roughly equal lengths as the real words. While the pseudo words were clearly non-existing, they obeyed English phonotactic constraints. We ensured that participants could not guess from which real words the pseudo words were derived by having four Dutch native listeners who did not participate in Experiments 1 and 2 indicate whether the words strongly reminded them of particular real English words. If so, we altered additional phonemes until this was no longer the case.

In the first part of the experiment, 30 experimental real words occurred as primes, in addition to their 30 counterpart pseudo words. The second part of the experiment contained repeats from the first part: the 30 experimental real words appeared as targets, in addition to their counterpart pseudo words. Additionally, the second part of the experiment contained 20 new distractor foils (i.e., these were not repeats from the first part), consisting of 10 real words and 10 counterpart pseudo words. Both parts of the experiment started with the same 6 practice trials, consisting of 3 real words and 3 counterpart pseudo words. The design of the experiment is summarized in Table 5.1.

The experimental real words had an average log-transformed frequency of occurrence of 4.48 per million (*SD* = 1.97); real word distractor foils of 4.21 per million (*SD* = 2.58), and real word practice items of 2.54 per million (*SD* = 2.26; British National Corpus, version 1.0, 1995). There is no statistical difference in the frequency of occurrence between the experimental real words and the real word distractor foils (*t* (13) = -0.29).

Table 5.1: Overview of stimulus types occurring in Parts 1 and 2 of Experiments 1 and 2 with examples.

Part 1	
60 pseudo-randomized stimuli, consisting of:	
- 30 experimental real words (primes)	*balloon*
- 30 counterpart pseudo word fillers	*ballee*
Part 2	
80 pseudo-randomized stimuli, consisting of:	
• 60 repetitions of Part 1, consisting of:	
- 30 experimental real words (targets)	*balloon*
- 30 counterpart pseudo words	*ballee*
• 20 new distractor foils, consisting of:	
- 10 real words	*result*
- 10 counterpart pseudo words	*rezell*

We recorded all real and pseudo words (i.e. experimental real words, real word distractor foils, real word practice items, and all of their counterpart pseudo words) with a male native speaker of British English, and we also recorded the experimental real words and their derived pseudo words (i.e., a subset of the materials) with a female native speaker of British English. The speakers read the real and pseudo words from paper in a sound-attenuating booth, and were recorded with a Sennheiser ME 64 microphone and Adobe Audition 1.5 recording software at a sampling rate of 44.1 kHz at 2 bytes/sample. We recorded multiple tokens for each real word and pseudo word with each speaker. Editing (e.g., cutting of the long audio file into individual stimulus wave files and amplitude equalizing) was done with Praat software (Boersma & Weenink, 2018). Depending on whether the real word or pseudo word was to occur once or twice in the experiment, we selected the one or two best sounding tokens from the male speaker. From the female speaker, we only selected the best sounding token for each real and pseudo word. The two tokens of the experimental real words produced by the male speaker had an average duration of 561 ms (*SD* = 72, range: 410-740), and the average duration of the tokens of the experimental real words produced by the female speaker was 642 ms (*SD* = 62, range: 551-831), a statistically significant difference ($\hat{\beta}_{\text{male speaker}}$ = -80.2, t = -5.2, p < .001). One of the differences between our speakers thus appears to be their natural speech rate.

We presented primes (occurring in the first part of the experiment) and targets (occurring in the second part) in either the same or a different voice (in the match and the mismatch conditions, respectively). A match meant that a prime and target were both uttered by the male speaker, while a mismatch meant that the prime was uttered by the female speaker and the target by the male speaker (also see Table 5.2). For the first part of the main experiment, we created four lists, in each of which half of the stimuli were uttered by the female speaker and half by the male speaker. These lists contained the experimental real words and their counterpart pseudo words. The lists had different pseudo-randomized stimulus orders, and differed in which primes were produced by which speaker. Maximally three real or pseudo words followed each other. For each of the four lists, we created a mirror list in which we replaced the tokens produced by one speaker by tokens produced by the other speaker. For the second part, we again created four pseudo-randomized lists, in which the maximal consecutive number of real or pseudo words was also three. These lists contained new tokens of the experimental words and their counterpart pseudo words (i.e., repetitions from part 1), as well as tokens for the distractor foil real and pseudo words. All stimuli in the second part were produced by the male speaker. We paired these lists with the lists for part 1; every participant heard one pair of lists. Both parts started with six practice trials. For the first part, three were produced by the female speaker and three by the male speaker, while for the second part, all six were produced by the male speaker. The order of presentation of the practice items differed per part, but was the same for all participants.

Table 5.2: Experimental conditions and examples of primes and targets in Experiments 1 and 2. 'Exp.' denotes experiment. Phonetic transcriptions are given for stimuli in Experiment 2, which could undergo speech reduction.

Condition	Prime	Target
Match		
Exp. 1: male speaker - male speaker	*balloon*$_{\text{male}}$	*balloon*$_{\text{male}}$
Exp. 2: reduced token - reduced token	*b*^{a}*lloon*$_{\text{reduced}}$ /b^{ə}ˈluːn/	*b*^{a}*lloon*$_{\text{reduced}}$ /b^{ə}ˈluːn/
Mismatch		
Exp. 1: female speaker - male speaker	*tomato*$_{\text{female}}$	*tomato*$_{\text{male}}$
Exp. 2: unreduced token - reduced token	*tomato*$_{\text{unreduced}}$ /təˈmɑːtəʊ/	*t*^{o}*mato*$_{\text{reduced}}$ /t^{ə}ˈmɑːtəʊ/

Procedure

Participants were tested individually in a sound-attenuating booth (the English listeners at Cambridge University in the U.K., the Dutch listeners at the Max Planck Institute for Psycholinguistics in Nijmegen, the Netherlands, and the Spanish listeners at the Escuela Técnica Superior de Ingenieros de Telecomunicación of the Universidad Politécnica de Madrid in Spain). We presented stimuli via closed headphones at a comfortable listening level using E-prime 2.0 software (Psychology Software Tools, Pittsburgh, PA). We instructed participants to decide as quickly and accurately as possible whether the stimulus they heard was a real English word or not. Participants responded by pressing *yes* with their dominant hand (key *m* on a keyboard for right-handers, *z* for left-handers) or *no* with their non-dominant hand (*z* for right-handers, *m* for left-handers). The two parts of the experiment directly followed each other; there was no break between them.

Each trial started with the presentation of a blank screen for 300 ms. A fixation asterisk then appeared in the middle of the screen for 250 ms, followed by the auditory presentation of the stimulus (blank screen). The next trial started after the response, or in case of no response after three seconds from word onset. We recorded participants' response times (RTs) and accuracies. This experiment took approximately 20 minutes.

For the two groups of non-native listeners, the main experiment was followed by the LexTALE task (Lemhöfer & Broersma, 2012) and a questionnaire. LexTALE is an unspeeded visual lexical decision task aimed at testing English proficiency. Although it tests for vocabulary knowledge, its results have been shown to correlate substantially with general proficiency (Lemhöfer & Broersma, 2012). In this task, stimuli (40 real English words, 20 pseudo words, and three practice items) are presented one by one in a pseudo-randomized order on a computer screen in a black font (Arial Unicode MS, point size 18) in the middle of a white background with E-prime 2.0 software. Participants responded via the keyboard, by pressing *yes* with their dominant hand (key *m* on a keyboard for right-handed, *z* for left-handed) and *no* with their non-dominant hand (*z* for right-handed, *m* for left-handed). The next trial appeared on the screen upon the response key press. After the LexTALE task, participants filled out a language background questionnaire, in which they described their experience with English, and in which they self-rated their proficiency levels. The LexTALE task and the questionnaire each took approximately five minutes.

Analyses

Prior to the analyses, we excluded participants and target words whose error rates were separated 2.5 *SD* or more from the average error rates for participants and target words, respectively. To reduce noise in the data, we also discarded targets whose primes did not receive correct responses. Specific to the RT analysis, we also excluded targets with incorrect responses, and targets with reaction times exceeding 2.5 *SD* from the grand mean. After running the statistical model on the RTs, we discarded reaction times for which the residual standard errors deviated more than 2.5 times from the values predicted by the statistical model of the RTs because these were considered outliers, after which we refitted the model.

We analyzed the accuracy of the responses to the targets in the auditory lexical decision task with logistic mixed effects regression models with the binomial link function, and we analyzed the log-transformed reaction times (log RTs) to targets with mixed effects regression models.

In both the log RT and accuracy analyses, we tested the influence of our predictors of interest *Speaker match* (which reflects whether a prime and target were uttered by the same or a different speaker) and *Listener group*, and the interaction between these. *Listener group* was Helmert-coded (e.g., Fox & Monette, 2002) to assess within the same statistical model if a) the two groups of non-native listeners differed from the native listeners (Contrast 1), and b) if the two non-native groups differed amongst each other (Contrast 2). To capture additional variance in our data, we tested for effects of a number of control predictors that have been shown to affect speech processing in similar experiments (e.g., Hanique et al., 2013; Morano et al., in press): log-transformed *Target duration*, log-transformed *Log word frequency* as obtained from the British National Corpus (version 1.0, 1995), *Trial number*, and *Lag* in terms of trials between prime and target. Specific to the analysis of the log RTs, we additionally tested for effects of the control predictors log-transformed reaction times to the prime (*RT prime*) and to the preceding trial (*RT preceding trial*).

We used *Word* and *Participant* as crossed random effects in the analyses of the log RTs and accuracy, and tested for random slopes for the predictors of interest and their interactions (*Speaker match* by-participant, *Listener Group* and *Speaker match* x *Listener Group* by-word). We did not test for random slopes of our control predictors for three reasons: first, we had no experimental hypotheses about those; second, doing so increases the chances of overfitting the models to this particular dataset (which decreases the generalizability of our findings); and third, doing so increases the chances of model convergence failures. We only included statistically significant effects and interactions in the model (i.e., whose absolute t-values exceeded 1.96, which implies p < .05 for this type of data with many observations that together approach a normal distribution), as well as simple effects of predictors appearing in significant interactions.

Results and discussion

According to the outlier criteria on error rates described above, we excluded two Spanish non-native listeners and one Dutch non-native listener as well as the target ‘saloon’ from the analyses.

Participants, on average, made 3% of errors on the targets (*SD* = 3%; native English listeners: 2%; Dutch listeners: 3%; Spanish listeners: 4%; *SD* = 4%, 3%, 3%, respectively). As we found no effect of *Listener group*, there was no statistically significant difference between the English listeners on the one hand and the Dutch and Spanish listeners on the other hand (i.e. Contrast 1), or between the Dutch and the Spanish listeners (i.e., Contrast 2). No simple effect of *Speaker match* arose, nor did we observe an interaction between *Speaker match* and *Listener group*.

RTs ranged from 558 ms to 1656 ms, and were 936 ms on average (*SD* = 186 ms). Our statistical model for the log-transformed RTs (see Table 5.3) shows effects of the control predictors *Log RT previous* and *Log RT prime*, indicating faster responses to targets whose primes or preceding trials received quick responses. We also observed an effect of Contrast 1 of *Listener group*, showing that the native English listeners responded significantly more quickly than the Dutch and Spanish listeners did.

More importantly for our research question, we observed an interaction between our predictors of interest *Speaker match* and *Listener group*. The statistical model

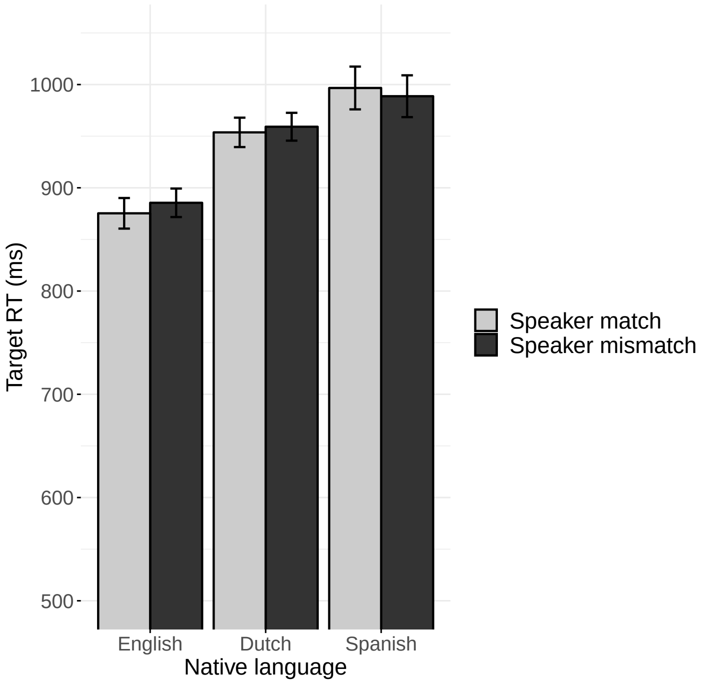
Table 5.3: Statistical model for RTs of correct responses to targets in Experiment 1. The intercept represents Speaker mismatch. Contrast 1 compares native (0.666) to non-native listeners (both -0.333), and Contrast 2 compares Dutch (0.5) to Spanish (-0.5) non-native listeners.

Fixed effects	$\hat{\beta}$	*t*
Intercept	5.14	41.4
Speaker match (match)	-0.003	-0.6
Log RT previous	0.06	5.6
Log RT prime	0.18	12.5
Contrast 1 (native vs. non-native)	-0.07	-4.2
Contrast 2 (Dutch vs. Spanish)	-0.004	-0.2
Speaker match (match) x Contrast 1	-0.02	-2.2
Speaker match (match) x Contrast 2	-0.01	-1.0
Random effects		*SD*
Word	intercept	0.05
Participant	intercept	0.07
Residual		0.13

shows that the effect of *Speaker match* differs between the native listeners and non-native listeners (Contrast 1), but not between the Dutch and Spanish listeners (Contrast 2; also see Figure 5.1). We ran statistical models without the simple and interaction effects of *Listener group* on the data split according to Contrast 1 to interpret the difference between the native and the non-native listeners. The model for the native listeners showed a significant effect of *Speaker match* ($\hat{\beta}_{\text{speaker match}}$ = -0.02; *t* = -2.2). In contrast, in the model applied to the data of the non-native listeners, *Speaker match* was not statistically significant ($\hat{\beta}_{\text{speaker match}}$ = -0.01; *t* = -1.1).

In summary, Experiment 1 showed a difference in exemplar effects between native and non-native listeners in the log RTs, indicating that larger exemplar effects arose for the native than the non-native listeners.

Figure 5.1: RTs of analyzed correct responses to targets in Experiment 1, split according to participants' native languages, and speaker match condition. Error bars represent 95% confidence intervals.



Experiment 2

Method

Participants

A total of 114 listeners (who did not participate in Experiment 1) took part in Experiment 2. Thirty-four were native English listeners (13 males, five left-handers, mean age: 22 years), 40 were native Dutch listeners (11 male, two left-handed, mean age: 20 years, mean LexTALE score: 73%, *SD* = 12%; mean self-assessed listening proficiency: 4.7, *SD* = 0.7), and 40 were native Spanish listeners (24 male, all right-handed, mean age: 22 years; mean LexTALE score: 67%, *SD* = 10%; mean self-assessed listening proficiency: 3.8, *SD* = 1.1). The two non-natives groups' English proficiencies can be considered as roughly equal, as both of their average LexTALE scores fall within the range of CEFR level B2.

Materials

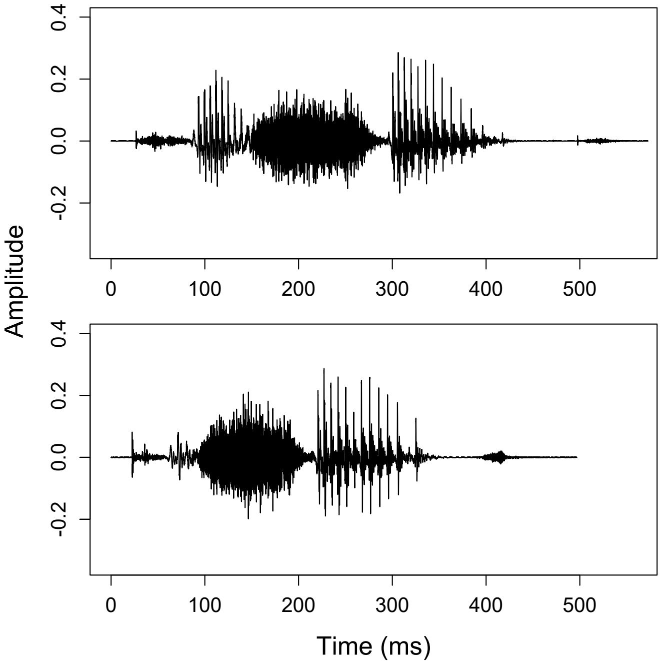
We used the same word types as in Experiment 1. In Experiment 1, stimuli were uttered in one of two speaker voices (male or female), and stimulus repetitions could be in the same or a different voice (male - male or female - male). In Experiment 2, stimuli are reduced or unreduced pronunciation variants, all produced by the same speaker. The speaker was the male speaker from Experiment 1. Our reduced pronunciation variants are characterized by shorter overall durations and shortened segments typical of a casual speech style; especially the vowel in the initial syllable was substantially shorter. An example of reduced and unreduced token of the experimental real word ‘cassette’ is given in Figure 5.2.

We re-used stimuli produced by the male speaker from Experiment 1 as unreduced stimuli in Experiment 2. For the reduced stimuli, we made new recordings with the same speaker and the same recording equipment. The speaker was instructed to produce tokens as if in casual speech. We selected the two best sounding tokens for to-be-repeated stimuli (i.e., experimental words, practice items and their respective counterpart pseudo words), and the single best sounding tokens for stimuli that were not to be repeated (i.e., the real word distractor foils and their counterpart pseudo words).

Unreduced tokens of the experimental real words were 628 ms on average (SD = 75 ms), while reduced tokens of the experimental real words were 528 ms on average (SD = 68 ms), a statistically significant difference ($\hat{\beta}_{\text{unreduced}} = 100.3$, $t = 6.3$, $p < .001$). The durations of the reduced and unreduced primes and reduced targets are illustrated in Figure 5.3. Vowels in the initial syllable of unreduced tokens were also statistically longer than the same vowels in the reduced tokens (means of 59 ms, SD = 10 versus 24 ms, SD = 16; $\hat{\beta}_{\text{unreduced}} = 34.8$, $t = 11.2$, $p < .001$). Even though all reduced word tokens had a small portion of the vowel left, the vowels in these tokens were perceptually close to absent.

Reductions (or a lack thereof) typically occur over wider contexts than single segments, as they are characteristic of an overall speech style. If our tokens are naturalistic in this respect, not only the vowel in the tokens’ initial syllables should be reduced (or unreduced), but also the rest of the tokens should show a certain (or no) degree of reduction. We tested whether this was the case by first subtracting the vowel duration from the total duration of the tokens of each experimental word (reduced and unreduced tokens). We then tested whether vowel duration could pre-

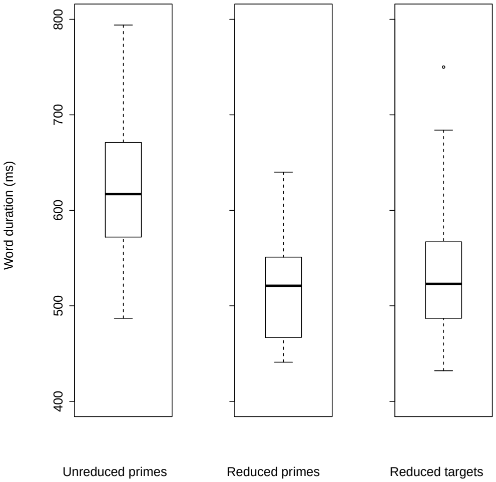
Figure 5.2: Examples of stimuli: an unreduced (top) and a reduced (bottom) token of the experimental real word 'cassette' /kə'sɛt/. The figure shows a substantial difference in the two tokens' overall duration as well as in the realization of the vowels in the two tokens' initial syllables.



dict remaining token duration. We found that this was the case ($\hat{\beta}_{\text{vowel duration}}$ = 0.97, t = 2.4, p < .05), which shows that our tokens are compatible with overall reduced or unreduced speech styles.

The experimental lists were identical to the ones used in Experiment 1, except that we adapted them to the new manipulation. In Experiment 1, primes were spoken by the male or female speaker (in the match and mismatch conditions, respectively), while all targets were spoken by the male speaker. In Experiment 2, primes were reduced or unreduced tokens produced by the male speaker, while all targets were reduced tokens produced by the male speaker (see also Table 5.2). We replaced all of the male speaker's tokens by reduced tokens, and replaced all tokens by the female speaker by unreduced tokens.

Figure 5.3: Distribution of the durations (in ms) of the unreduced primes (left panel), the reduced primes (middle panel) and the reduced targets (right panel) in Experiment 2.



Procedure and analyses

The procedure and analyses were identical to the ones in Experiment 1, except that the predictor of interest *Speaker match* in our statistical analysis was now *Variant match*.

Results and discussion

According to our 2.5 *SD* outlier exclusion criterion for errors, we excluded three Spanish non-native listeners, and, as in Experiment 1, we discarded the word ‘saloon’ from the analyses.

Participants made, on average, 4% of errors (*SD* = 4%) on the targets (native English speakers: 3%; Dutch listeners: 4%; Spanish listeners: 6%; *SD* = 3%; 4%; 4%; respectively). The errors showed an effect of *Listener group*, indicating that the English natives made fewer mistakes than the Dutch and Spanish listeners (Contrast 1, $\hat{\beta}$ = 0.63, t = 2.3, p < .05). There was no difference between the Dutch and the Spanish listeners (Contrast 2, $\hat{\beta}$ = 0.44, t = 1.6, p = .1). No simple effect

Table 5.4: Statistical model for log RTs of correct responses to targets in Experiment 2. The intercept represents *Variant mismatch*. Contrast 1 compares native (0.666) to non-native listeners (both -0.333), and Contrast 2 Dutch (0.5) to Spanish (-0.5) non-native listeners.

Fixed effects	$\hat{\beta}$	*t*
Intercept	5.65	56.2
Variant match (match)	-0.01	-2.0
Log RT prime	0.17	11.6
Contrast 1 (native vs. non-native)	-0.10	-4.7
Contrast 2 (Dutch vs. Spanish)	-0.10	-4.4
Variant match (match) x Contrast 1	0.04	3.2
Variant match (match) x Contrast 2	0.03	2.2
Random effects		*SD*
Word	intercept	0.05
Participant	intercept	0.09
Residual		0.14

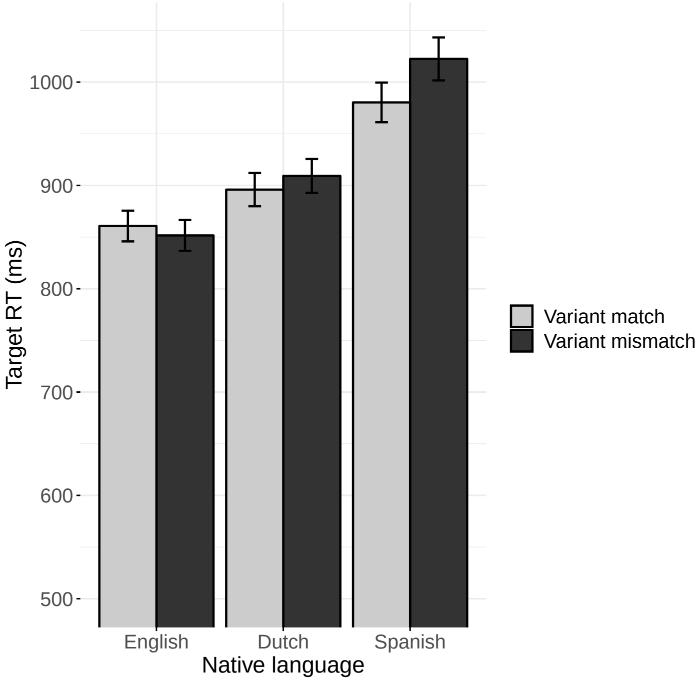
of *Variant match* arose, nor did we find an interaction between *Variant match* and *Listener group*.

RTs ranged from 536 ms to 1679 ms, and were 917 ms on average (*SD* = 205 ms). Our statistical model to the log RTs, summarized in Table 5.4, shows that *Log RT prime* is a highly significant predictor for log RTs, indicating that responses were significantly quicker to targets whose primes received quick responses.

More importantly for our research question, we obtained a statistically significant interaction between *Variant match* and *Listener group*, showing that the two groups of non-native listeners differed from the English native listeners in *Variant match*. In addition, the interaction showed that the Dutch and Spanish listeners differed from each other in *Variant match*. Figure 5.4 illustrates the *Variant match* effects for all groups.

To interpret the *Listener group* x *Variant match* interaction, we ran models with the same predictors as before, apart from the simple effect of *Listener group* and its interaction term, on subsets of the native and non-native listener data (i.e., to interpret Contrast 1). In these models, we found significant exemplar effects for the non-native listeners ($\hat{\beta}_{\text{variant match}}$ = -0.02; *t* = -3.0), but not the native listeners

Figure 5.4: RTs of correct responses to targets in Experiment 2, split according to participants' native languages, and variant match condition. Error bars represent 95% confidence intervals.



($\hat{\beta}_{\text{variant match}}$ = 0.01; t = 1.4).

In addition, we ran models on the separate data of the Dutch and the Spanish listeners (i.e., to interpret Contrast 2). In these models, we observed significant exemplar effects for the Spanish listeners ($\hat{\beta}_{\text{variant match}}$ = -0.03; t = -3.1), but not for the Dutch listeners ($\hat{\beta}_{\text{variant match}}$ = -0.01; t = -1.0).

In summary, we established a difference in exemplar effects between our native and non-native listener groups. Unlike in Experiment 1, where larger exemplar effects arose for the native listeners, in Experiment 2, the exemplar effects held for the non-native listeners only. Also at odds with Experiment 1 (where exemplar effects did not differ statistically between the Dutch and Spanish listeners), we obtained a difference in exemplar effects between the two groups of non-native listeners, showing that exemplar effects were larger for the Spanish listeners than for the Dutch listeners.

General discussion

In the present study, we investigated the occurrence and size of exemplar effects in native and non-native listening. Our first research question was whether the size of exemplar effects varies between native and non-native listening, which would show that native and non-native listening differ from each other in a qualitative way. We addressed this question by comparing exemplar effects in native and non-native listener populations in Experiments 1 and 2. Our second research question was how non-native listeners' L1 phonology interacts with the size or occurrence of exemplar effects. We investigated this question by comparing exemplar effects of a largely language-specific variation type (vowel reduction in English) between two non-native populations who differ in their familiarity with the variation type from their L1 phonology in Experiment 2.

In both experiments, we tested native English listeners as well as Dutch and Spanish non-native listeners of English. The two groups of non-native listeners were at roughly the same proficiency level in English: CEFR level B2, corresponding to 'upper intermediate' proficiency. In Experiment 1, as in most preceding literature, the variation type that formed the basis of the match/mismatch condition was speaker voice. All listeners have ample experience dealing with this indexical variation type. For this experiment, we found that exemplar effects were larger for native than for non-native listeners. We did not observe statistically significant differences between the Dutch and the Spanish non-native listeners.

The difference in exemplar effects between native and non-native listeners speaks to our first research question. Importantly, our study is the first to clearly establish such a difference, as the two previous studies that compared native and non-native listening did not report or found no statistically significant differences (Trofimovich, 2005; Winters et al., 2013). Crucially, our finding may be indicative of qualitative differences between native and non-native speech perception.

The exemplar effects that differed between native and non-native listeners in Experiment 1 may be accounted for in terms of differential availability of processing resources. A lexical decision task places different cognitive demands on native and non-native listeners: while it is quite an easy task for native listeners, this is not at all true for non-native listeners. Non-natives' weaker and less well-specified lexical representations (e.g., Cook, Pandža, Lancaster, & Gor, 2016) introduce a considerable amount of uncertainty, which makes it harder for these listeners to accept or reject stimuli as real words. Moreover, the fact that a given word is not present in a non-native listener's lexicon does not necessarily mean the word is not real. As a result of this enhanced task difficulty, non-native listeners likely have few processing resources available to engage in other processes than those directly necessary for consulting the mental lexicon (other processes such as the reliance on exemplars). We believe that the non-native listeners in our experiment did consequently not exploit the prime-target voice matching to the same extent as the native listeners did. Apparently, when the processing load in a lexical decision task is relatively high, listeners rather rely on abstract representations than on exemplars.

The lack of exemplar effects for the non-native listeners in Experiment 1 may appear at odds with the earlier studies which established exemplar effects for speaker voice for L2 listeners. The task in these two studies, however, was not lexical decision as in our study. In Drozdova et al. (2019) and Winters et al. (2013), exemplar effects arose in an old-new judgment task, which does not involve a costly lexical search. In fact, one of the listener groups in the Winters study was able to perform the old-new judgment task without knowledge of the target language, again showing that this task could be performed with little to no lexical involvement. Importantly, in the word identification task of Drozdova et al., a task which involves lexical access like our lexical decision task does, exemplar effects did not arise, as in our experiment. Trofimovich (2005)'s task was also very different from ours: listeners had to repeat out loud spoken prompts. This task thus involved listeners' production of words. In addition, and perhaps more importantly, like in the Winters study, this task can be performed with significantly less involvement of the mental lexicon than a lexical decision task. Task differences thus likely explain the discrepancy between Experiment 1 and the studies by Drozdova et al., Trofimovich, and Winters et al.

In Experiment 2, primes and targets were produced by the same speaker, but could match or mismatch in their degree of speech reduction. Reduced word tokens had highly shortened vowels in their initial, unstressed syllables and shorter overall durations than unreduced word tokens. This reduction occurs frequently in English (e.g., Shockey, 2003), and is therefore highly familiar to native listeners. Native listeners process reduced word tokens with ease (especially so in running speech, e.g., Ernestus, Baayen, & Schreuder, 2002). In contrast, non-native listeners lack exposure to reduced word tokens and are therefore less familiar with such tokens, which results in perceptual difficulties (e.g., Ernestus, Dikmans, & Giezenaar, 2017; Henrichsen, 1984). Importantly, there may be a difference between Dutch and Spanish listeners in how easily they process the reduced word tokens in the experiment. Dutch has a similar vowel reduction pattern as the one in our study, which offers the Dutch listeners familiarity with the reduction pattern through their L1 phonology. Spanish, on the other hand, does not reduce its vowels, nor does its phonology include the schwa vowel. The Spanish non-native listeners are therefore neither familiar with the vowel in the initial syllable of most of the experimental words, nor with its reduction through their L1 phonology.

As in Experiment 1, and of relevance for our first research question, we obtained a significant difference in exemplar effects between the native and the non-native listeners. This finding again indicates that the non-native listeners showed qualitative differences in processing relative to the native listeners.

At odds with Experiment 1, exemplar effects were larger for the non-native listeners than for the native listeners in Experiment 2. When we examined the data of the native and non-native listener groups separately, we found that the non-native listeners showed statistically significant exemplar effects, while the native listeners showed no significant exemplar effects. We also found a significant difference in exemplar effects between the Dutch and the Spanish listeners: larger exemplar effects arose for the Spanish listeners than for the Dutch listeners.

In Experiment 1, we ascribed the difference in exemplar effects between the native and non-native listeners (whereby exemplar effects only arose for the native listeners) to a differential availability of processing resources for the two listener groups. As we observed very different results in Experiment 2 (i.e., exemplar effects only for the non-native listeners), it appears that the largely language-specific variation type introduced in Experiment 2 (an experiment which was otherwise identical to Experiment 1) was responsible for this different pattern of exemplar effects between the native and non-native listener groups.

All targets in Experiment 2 were reduced tokens, and half of the primes were too. For non-native listeners, the recognition of such tokens is particularly challenging. The fact that many words appeared as reduced tokens twice in the experiment is a factor that likely invited the non-native listeners to consult the exemplars created by the primes to recognize the reduced targets, as doing so may aid the recognition of these tokens. Thus, while non-native listeners appear to use abstract representations in tasks with a high processing load (such as the lexical decision task in our study), they appear to only do so when the variation in the experiment is known (such as the speaker voice variation in Experiment 1). Instead, when variation is unknown or less familiar, such as the reductions were to the non-native listeners in Experiment 2, reliance on exemplars may offer a processing benefit in a context where words are repeated. As the task was identical between experiments (and thus required similar amounts of processing resources), this appears to be the case also when the experimental task requires many processing resources.

The difference we observed between the Dutch and the Spanish listeners in exemplar effects in Experiment 2 is in line with the account above (i.e., reliance on exemplars as a result of low familiarity with a variation type). The Spanish listeners are most unfamiliar with the reduced tokens that occurred in Experiment 2, because Spanish does not have schwa in its phonology, nor does it reduce its vowels. The Dutch listeners are more familiar with the reduced tokens since Dutch listeners are used to vowel reduction from their L1 phonology. We observed larger exemplar effects for the Spanish listeners, the listener group with the smallest familiarity with the reductions. Especially this group of listeners relied on exemplars, probably because doing so aided these listeners the most in the recognition of reduced targets.

As anticipated, the two variation types in our experiments (speaker voice and vowel reduction) exerted different influences on the non-native listeners in our experiments. We attributed these differences to non-native listeners' varying familiarity with the two variation types. A finding that was less expected is that the native listeners behaved differently in Experiments 1 and 2 as well. The native listeners are perfectly familiar with both speaker voice variation and vowel reduction, and may therefore show similar exemplar effects for either of these variation types. Instead, we observed that statistically significant exemplar effects arose for the native listeners in Experiment 1, while this was not the case in Experiment 2. Possibly, this discrepancy between experiments arose due to differences in salience of the variation types to the native listeners. The variation in speaker voice as used in Experiment 1 is extremely obvious to listeners (whether they be native or non-native). Instead, the within-speaker variation in reduction used in Experiment 2 is much more subtle to native listeners (note that this stands in sharp contrast with non-native listeners, for whom speech reductions have drastic processing consequences). As vowel reductions are extremely common in English, they are not likely to stand out to native listeners. Therefore, we hypothesize that, likely having sufficient processing resources available in either experiment to show exemplar effects, the native listeners only showed them in Experiment 1 because only in that experiment they clearly noticed the prime-target variation (in speaker voice).

The exemplar effects in non-native listening extend the exemplar effects that arose for speaker voice in non-native listening in Winters et al. (2013) and Trofimovich (2005). In addition, they add to the exemplar effects in non-native listening for language-specific L2 variation established in one of the experiment versions of Morano et al. (in press). Our study additionally reveals that multiple factors are at play in determining non-native listeners' use of exemplars. Future research should further investigate the role of exemplar effects in non-native listening. Meanwhile, in order to accommodate the findings so far, theories of non-native language listening should be adapted in order to incorporate exemplars in addition to abstract lexical representations.

In sum, we found that both native and non-native listeners may show exemplar effects in auditory identity priming experiments. Our study showed that the occurrence of exemplar effects for these listeners is mediated by at least three factors: first, listeners' availability of processing resources (whereby exemplar effects are more likely to arise when listeners have more processing resources available); second, their familiarity from the L1 phonology with the variation type that forms the basis of the match/mismatch condition (whereby an L1-imposed lack of experience with a variation type may lead to listeners' increased reliance on exemplars); and third, the extent to which variation stands out to listeners, a factor also influenced by listeners' L1 experience (whereby exemplar effects are more likely to arise for more salient variation types). Most importantly, both experiments clearly showed that the occurrence of exemplar effects differs for native and non-native listeners. Crucially, this suggests that the type of representations (abstract representations or exemplars) primarily used in speech comprehension varies between native and non-native listening, which points at qualitative differences between native and non-native listening.

General discussion and conclusions

Chapter 6

Models of speech comprehension differ in their assumptions about how the pronunciation of words are stored. Abstractionist models (e.g., McClelland & Elman, 1986; Norris, 1994) pose that a typical word has a single lexical representation, which consist of a sequence of abstract units such as phonemes. Situation and speaker-specific information of tokens is not stored in abstract lexical representations. Exemplar models (e.g., Goldinger, 1998; Johnson, 1997), instead, assume that a word is stored as many individual tokens, which together form a cloud for that word. These tokens are represented in full phonetic detail, and thus retain situation- and speaker-specific information. Hybrid models (e.g., Goldinger, 2007; McLennan et al., 2003) assume that abstract representations and exemplars coexist.

This dissertation investigated the nature of exemplars and their role in speech comprehension by testing the hypothesis that exemplars are represented in episodic memory rather than in the mental lexicon. It presents a series of studies to address two subquestions. The first subquestion concerns the circumstances under which exemplars affect speech perception, and the second subquestion addresses the types of surface variation that exemplars may contain. This chapter presents the results of these studies, describes how they improve our understanding of exemplars, and provides recommendations for future research.

Chapters 2 and 3

Chapters 2 and 3 focused on the circumstances under which exemplar effects arise in auditory identity priming experiments. Chapter 2 did so by reviewing the literature on exemplar effects, and assessed to what extent findings in the literature are in accordance with the hypothesis that exemplars are represented in episodic memory. We formulated four predictions on when exemplar effects should be largest in auditory identity priming experiments if exemplars were to be represented in episodic memory. The first prediction stated that exemplar effects are larger in experiments with tasks that instruct participants to use their episodic memory than in experiments with tasks that do not. This prediction was supported by one study only, and therefore requires additional research. The second prediction stated that exemplar effects are larger in experiments with less intelligible stimuli, because such stimuli invite participants to rely on their episodic memories. This prediction was validated by several previous findings. The third prediction stated that exemplar effects are larger for primes that are less consistent with prior knowledge (such as low frequency words), because such primes integrate less well with prior knowledge, making them more likely to still be available to participants when they process the targets. This prediction also received support from several studies. The fourth prediction stated that exemplar effects are larger for primes that are processed for perceptual properties (for instance, in a task focusing participants' attention on primes' loudness) because such perceptual processing slows down memory integration. The literature was indecisive about this prediction because the relevant studies may have been confounded. None of the predictions was thus contradicted by previous findings. On the basis of the literature, it is likely that exemplars are not represented in the mental lexicon, but in episodic memory. It was also clear, however, that more research was necessary to strengthen this hypothesis.

Chapter 2 also investigated the hypothesis that exemplars are represented in episodic memory rather than in the mental lexicon more directly in two long-term identity priming experiments. The experiments tested whether exemplar effects are larger if participants' task on the primes requires processing of perceptual (categorization of primes' loudness or speaker voice) rather than meaning-related properties (categorization of whether primes refer to objects typically found indoors or outdoors) of primes. If so, this would be consistent with episodic memory as the locus for exemplars. No effect of the task on the primes was found (i.e., exemplar effects arose independently of prime task), and it was therefore not possible to link the occurrence of exemplar effects to properties of episodic memory.

Chapter 2 also tested whether exemplar effects arise more reliably when the task on targets in an experiment instructs participants to use their episodic memories than when the task does not instruct them to do so. If so, this would suggest that exemplars are represented in episodic memory. Chapter 3 focused on the same question. In both of these studies, the occurrence and size of exemplar effects in an old-new judgment task and in a semantic classification task on the targets was compared. The old-new judgment task asks participants to categorize stimuli according to whether or not they occurred previously in the experiment, which requires participants specifically to consult their episodic memories. The semantic classification task (in which participants judged the animacy of referents), in contrast, does not require them to do so. Chapter 2 collected participants' reaction times (RTs) and response accuracy. For the old-new judgment task, clear exemplar effects arose in the RTs. In the response accuracy data, exemplar effects were less robust, as they were restricted to only one of the talkers who produced the stimuli in the experiment. For the semantic classification task, participants appeared to rely on abstract representations rather than on exemplars. Chapter 3 collected participants' RTs, response accuracy, and EEG (electroencephalography). While no exemplar effects arose for either of the tasks in RTs and response accuracy, the ERP data replicated the exemplar effects of Chapter 2 for the old-new judgment task. These exemplar effects arose most clearly in early time windows of the EEG signal, which capture ERP (event-related potentials: stimulus-locked brain potentials derived from the EEG signal) components related to acoustic processing and lexical activation. The ERP findings for the semantic classification task were reversed from those for the old-new judgment task. The ERP data for the semantic classification task can be interpreted in different ways, but importantly for the current purposes, it is clear that this task yielded exemplar effects less clearly than the old-new judgment task did.

In both Chapters 2 and 3, exemplar effects for the old-new judgment task did not arise across the board because, for the response accuracy in Chapter 2, they were restricted to one of the two talkers in the experiment and, in Chapter 3, to the EEG data. Nevertheless, the two chapters together strongly suggest that exemplar effects are more likely to arise in an old-new judgment task than in a semantic classification task. This finding supports the hypothesis that exemplars are represented in episodic memory rather than in the mental lexicon.

Chapters 4 and 5

In Chapters 4 and 5, participants were invited to rely on their episodic memories to varying extents as a result of the level of listening difficulty in the experiment. When listening difficulty is enhanced, reliance on episodic memory offers participants in priming experiments a processing advantage. Difficult listening conditions therefore invite participants to rely on their episodic memories, while easier listening conditions do so to a lesser extent. If exemplars are represented in episodic memory, we expect exemplar effects to arise especially when participants are invited to use their episodic memories.

Chapter 4 manipulated listening difficulty by using background noise in a lexical decision experiment. In one experiment, participants listened to primes and targets presented in clean, while in another experiment, the same primes and targets were embedded in speech-shaped noise. Listening difficulty was thus enhanced in one experiment, while the noise level (i.c., the SNR) was chosen such that it was still possible to perform the task. Exemplar effects arose in the experiment with noisified stimuli, while exemplar effects were absent in the experiment with clean speech. Exemplar effects thus arose if participants were encouraged to use their episodic memories. Again, this finding is in line with the hypothesis that exemplars are represented in episodic memory.

The findings of Chapter 4 challenge the ‘time-course hypothesis’ of McLennan and Luce (2005). This hypothesis states that exemplar effects for indexical variation arise in late processing stages, because exemplars take time to become activated. Exemplar effects in our study only arose in the experiment with noisified stimuli. Crucially, responses were *faster* in that experiment (a somewhat surprising finding, which we attributed to participants’ increased motivation). In addition, exemplar effects arose exclusively for targets whose primes received fast responses. As in previous studies (e.g., Hanique et al., 2013), there was a strong relationship between reaction times on primes and targets (i.e., RTs on the primes could reliably predict RTs on targets). Indirectly, the restriction of exemplar effects to targets with fast primes thus suggests that exemplar effects only arose when participants responded quickly to targets. A final piece of evidence clearly speaking against the time-course hypothesis comes from Chapter 3 described above, in which exemplar effects arose most clearly early in the EEG signal. Importantly, in contrast to the time-course hypothesis, the hypothesis that exemplars are represented in episodic memory can account for exemplar effects in both early (as in Chapters 3 and 4) and late (as in the experiment with a difficult lexical decision task by McLennan & Luce, 2005) processing stages.

Chapter 5 varied listening difficulty by testing listener populations which have different levels of processing load. It presents two lexical decision experiments conducted with native listeners (L1) as well as two populations of non-native (L2) listeners. Compared to native listening, non-native listening is associated with a substantially higher processing load, as is evident from non-native listeners’ increased pupil response (a measure of cognitive effort), for instance (Borghini & Hazan, 2018). The higher processing load for the non-native listeners was expected to invite these listeners to rely on their episodic memories, and consequently, evoke exemplar effects for these listeners. However, it was unknown whether exemplars operate similarly in L1 and L2. In fact, previous research has hardly addressed whether exemplar effects ever arise in L2 listening.

Chapter 5 also addressed the question which extra-linguistic information types are stored in exemplars. If exemplars are based in episodic memory, no significant differences in exemplar effects are expected for different variation types, as all variation types can be stored in episodic memory. The mental lexicon, in contrast, likely has constraints on the variation types that are stored. One variation type used in Chapter 5, variation from speaker voice, is not language-specific. All listeners are therefore used to dealing with this variation type. The other variation type in the chapter stems from initial vowel reduction in English. This variation type, in contrast to speaker voice, is language-specific. Importantly, one group of non-native listeners was familiar with this type of reduction from their native language phonology (Dutch listeners), while the other group of non-native listeners was not (Spanish listeners). The experiments thus investigated whether there is a difference in exemplar effects between language-specific and language-unspecific variation types. In addition, in case of a language-specific variation type, the potential effect of being familiar with the variation type from the L1 phonology could emerge from the experiments.

Crucially, the non-native listeners showed exemplar effects for the language-specific variation type (vowel reduction). This finding contributes to the studies that showed that a range of variation types can lead to exemplar effects, and hints that exemplars are not based in the mental lexicon. In contrast, this result points to episodic memory as the locus for exemplars. Of interest, exemplar effects were larger as a function of how unfamiliar the listener groups were with the language-specific variation type (vowel reduction), suggesting that reliance on exemplars helps in processing words with unknown surface variation. The fact that the non-native listeners showed larger exemplar effects in the experiment with an unknown language-specific variation type (vowel reduction) than in the experiment with a known language-unspecific variation type (speaker voice) corroborates this idea.

Chapter 5 revealed two additional factors which mediate the occurrence of exemplar effects. The first factor is the amount of processing resources listeners have available: when more processing resources are available, exemplar effects are more likely to arise. This factor appeared from the experiment with speaker voice variation, in which the native listeners (who have many processing resources available) showed exemplar effects, while the non-native listeners (who have few processing resources available) did not. This factor may appear at odds with Prediction II from Chapter 2, which states that exemplar effects are larger for less intelligible targets. Note that the findings discussed in the context of that prediction were all obtained with native listeners, for whom the speech signal was made relatively more difficult to process. Other than the non-native listeners in the experiment with speaker voice variation in Chapter 5 however, these native listeners in principle had sufficient processing resources to perform the tasks.

The second factor is the extent to which a familiar variation type stands out to listeners: the more it stands out, the more likely exemplar effects are to arise. This factor was evident from the finding that native listeners showed exemplar effects for speaker voice variation (which was highly obvious to them), but not for within-speaker variation stemming from vowel reduction (which is less likely to stand out to native listeners, as vowel reductions are extremely common in English). The latter factor is related to attention. Based on this finding, it is possible that exemplar effects were not absent for amplitude variation in Palmeri et al. (1993) because amplitude is not a phonetically relevant variation type (compatible with the 'phonetic relevance hypothesis' by Sommers & Barcroft, 2006), but instead because the variation in amplitude did not particularly stand out to participants in that experiment.

Chapter 5 thus suggests that attention may play a role in the occurrence of exemplar effects. Interestingly, the reliance on exemplars therefore does not appear to be an automatic process. This finding is somewhat reminiscent of a study by Tuft, McLennan, and Krestar (2016), who observed larger exemplar effects when participants' general attention levels were increased as a result of the inclusion of taboo words in the experiments. Possibly, when a variation type is highly salient, participants are invited to exploit prime-target matching. A role for participants' attention to the variation type is also compatible with studies showing that exemplar effects may differ depending on whether the task on the primes directs participants' attention to the variation forming the basis of the match/mismatch condition (e.g., Theodore et al., 2015).

To get a better grip on the circumstances under which exemplars affect speech comprehension, future research should further investigate the role of attention on listeners' reliance on exemplars. It should be established how a variation type's salience and participants' resulting attention to prime-target matching may precisely affect the size or occurrence of exemplar effects. For instance, a dedicated experiment could test for exemplar effects of variation types with different salience levels (determined through rating studies, for instance). Other relevant questions are if reliance on exemplars is indeed not automatic, to what extent it is strategic, or even conscious. One way to tackle these issues would be to question participants about the listening strategies they used in the experiment in debriefings (for instance, whether they thought the repetition of words in the experiment was helpful in performing the experiment).

General implications

Together, the studies in this dissertation offer different pieces of evidence to suggest that exemplars are based in episodic memory rather than in the mental lexicon. First, it was established that the literature was mostly in line with episodic memory as the locus for exemplars. As such, a number of studies in which exemplar effects appeared to come and go in a seemingly unpredictable way (as those listed in 1.1) can now be accounted for. For instance, the exemplar effects that arose for word produced in a-typical manner in Nygaard et al. (2000) may have arisen because such words are more likely to remain available in episodic memory (because they are less easily integrated with lexical memory).

Second, participants' reliance on episodic memory – as a result of the task or as a result of listening circumstances – could be related to the occurrence of exemplar effects in experiments. Third, the experiments showed that exemplar effects may arise for variation in listeners' L2 unfamiliar from their L1 phonology, which also points at episodic memory as the locus for exemplars.

Another important finding is that exemplar effects differed significantly between native and non-native listening in both of the experiments reported in Chapter 5 (in the experiment with speaker voice variation, exemplar effects were significantly larger for the native listeners, and in the experiment with vowel reduction, exemplar effects were significantly larger for the non-native listeners). Chapter 5 is the first study to show that exemplars play different roles in native and non-native speech comprehension, as the two previous studies that compared native and non-native listening did not report or found no statistically significant differences between native and non-native listening (Winters et al., 2013; Trofimovich, 2005). Our finding hints at qualitative differences between the two types of speech comprehension processes.

Within the studies in this dissertation, exemplar effects clearly arose under some experimental conditions, while it was clear from interactions in the statistical analyses that under other experimental conditions within the same study, exemplar effects were significantly smaller or absent. There were thus also circumstances under which listening appeared to involve only or mostly listeners' use of abstract representations. As such, the findings in this dissertation neither support purely exemplar-based models, nor do they support purely abstractionist models.

An important question for future research to further our understanding of the speech comprehension process is how exemplars (based in episodic memory) interact with abstract representations (based in the mental lexicon) during on-line speech comprehension. For instance, do listeners activate abstract representations under all circumstances, and use exemplars *in addition to* abstract representations under specific circumstances? When they activate both abstract representations and exemplars, how do they link the information from the two?

Psycholinguistic experiments conducted in the highly controlled setting of a phonetics laboratory bear little resemblance to real life speech perception conditions. It is therefore the question to what extent results obtained in that setting generalize to more natural listening situations (see also Tucker & Ernestus, 2016). The experiments reported in this dissertation used somewhat more naturalistic speech materials than many previous studies did to improve the ecological validity of the experiments' findings compared to previous studies. In the match condition of almost all preceding studies, participants were twice presented with exactly the same recording (i.e., a frozen copy). Such identical tokens do not occur in real life, as two productions of the same word always differ, even when they are uttered consecutively by the same talker. In all of the experiments in this dissertation, different tokens for primes and targets were used (also in the match conditions). Moreover, ecological validity was enhanced by using 'normal' talkers without extensive phonetic training to produce the stimuli for the experiments. Despite the use of different tokens and despite the fact that untrained talkers produced the stimuli, exemplar effects arose. Crucially, this shows that exemplar effects generalize to more naturalistic listening conditions: they are not a mere by-product of repeating frozen copies, nor do they depend on whether stimuli are produced by phonetically trained talkers.

However, exemplar effects did not arise for all talkers that produced the stimuli. In Chapter 2, stimuli were produced by two talkers, and exemplar effects in response accuracy were restricted to the talker whose prime and target tokens sounded more similar. Thus, exemplar effects only appear when the natural item-to-item variation within prime-target pairs is small. All tokens for the experiments were produced as isolated words read from a list, which means that the item-to-item variation in the experiments in this dissertation was much smaller than the item-to-item variation of tokens occurring in natural conversations. Tokens in natural conversations occur with different intonations, with different speech rates, with different speech styles, in different sentence positions, etc. It is therefore the question whether exemplars play a substantial role in everyday speech comprehension, where much more token variation is present than in the experiments in this dissertation.

In addition, under normal, non-challenging listening conditions in which words are not frequently repeated, listeners likely do not rely on their episodic memories to a large extent. The observation that exemplar effects arise mostly when participants use their episodic memories therefore also suggests that the role of exemplars in everyday speech comprehension is likely small.

Methodological recommendations

Comparability between findings

The studies reported in this dissertation were carefully designed to test new manipulations, participant populations, and experimental measures while remaining comparable to previous studies, and among each other. For instance, to ensure comparability, all experiments reported in this dissertation used the same auditory long-term repetition priming paradigm, and most tested for exemplar effects of speaker voice variation. Within studies, we altered only the aspect under investigation while keeping all else strictly identical.

The design of Chapter 4 was directly based on one of the experiments reported in Hanique et al. (2013). In that experiment, exemplar effects arose, and the authors attributed these exemplar effects to the fact that word repetition was very obvious in their experiment, which invited participants' reliance on episodic memory. The exemplar effects in that study may, however, also have been driven by the presence of pronunciation variants without schwa in their experiment. Variants without schwa occur frequently in casual speech, but seldom out of context as in Hanique et al.'s study. These variants presented in isolation increased participants' processing cost, who were thus stimulated to use their episodic memories. This hypothesis was tested in an experiment reported in Chapter 4 with the same words that were produced with their schwa by two different talkers instead of by one talker with two pronunciation variants. If the exemplar effects of Hanique et al. were mostly due to the presence of reduced pronunciation variants in their experiment, they should disappear in the new experiment in which no reduced pronunciation variants occurred. Chapter 4 showed that exemplar effects were absent in the new experiment, suggesting that the exemplar effects of Hanique et al. were indeed to a large extent caused by the presence of reduced pronunciations variants. In addition, as exemplar effects reappeared in Chapter 4 when the same experiment was conducted with noise superimposed on the stimuli, the two studies combined show that listening difficulties stemming from different sources may lead to exemplar effects. A lesson for future research is that studies with designs based on previous studies may offer new insights on existing findings, and allow for interesting comparisons between existing and current findings.

In the analysis of the EEG data in Chapter 3, decisions had to be made up front about which time windows and electrode sites to include in the analysis. Such choices are usually informed by existing literature. Two possible problems are associated with this approach. First, if the research topic is relatively understudied, the literature may not provide clear guidelines as to which electrodes and time windows to examine. Second, ERP components may manifest differently under different circumstances (Luck, 2005). Only a handful of previous studies obtained exemplar effects in ERPs. The designs of these studies were not always well comparable to our study (for instance, because they used cross-modal priming, as in Friedrich, Kotz, Friederici, & Alter, 2004), and their results were somewhat mixed. Our study used electrode windows and electrode sites according to the clearest results in the literature, and found significant effects in two of our three analysis windows (the two early time windows). Importantly, the significant findings in Chapter 3 contribute to the small body of ERP studies that found exemplar effects in EEG, and will help guide analysis choices in future research. Specifically, I advise researchers planning ERP studies investigating exemplar effects to test for exemplar effects in early time windows over posterior electrode sites.

Mismatch effects

Before conducting the experiments in this dissertation, the expectation was that exemplar effects would arise in some experimental conditions, while smaller or no exemplar effects were expected to arise in other experimental conditions. Unexpectedly, *reversed* effects arose in some cases. In Chapter 2, participants' responses in the semantic classification task on the targets were *faster* in the mismatch condition for one of the talkers. Similarly, in Chapter 3, the expected amplitude differences in the ERPs between the match and the mismatch conditions in several windows were attested for the old-new judgment task, whereas for the semantic classification task, the pattern was reversed.

How to interpret mismatch effects is not obvious, also because until recently, the literature did not report any (a recent other example was documented by Morano et al., in press). It is unlikely that mismatch effects stem from the activation of exemplars, as more dissimilar prime exemplars should not enhance priming of their targets relative to more similar prime exemplars. Therefore, these effects are probably the result of another mechanism. In the semantic classification task in Chapter 2, the match condition consisted of primes and targets produced by the same talker, and the mismatch condition of primes by one talker and targets by the other talker. Participants responded faster in each experimental condition (i.e., both the speaker match and the mismatch conditions) in which primes were produced by the male talker rather than the female talker. As this facilitation arose independently of which talker produced the target, it must have reflected priming of the same abstract representation rather than be an exemplar effect. A rating study showed that primes by the male talker were more intelligible than those by the female talker. Probably, the more intelligible primes of the male talker activated word forms well, and therefore facilitated the recognition of the target to a large extent. The mismatch effect (with primes produced by the male talker and targets by the female talker) was thus likely due to the high intelligibility of the primes produced by the male talker. If so, the fast responses in the match condition (with primes and targets both produced by the male talker) were probably also caused by good prime intelligibility. Importantly, at first sight, this effect appeared to be a regular exemplar effect – it was only apparent through the reversed effect for targets produced by the female talker that the targets of both talkers probably just showed priming of abstract representations.

Similar to the interpretation of the mismatch effect in Chapter 2, Morano et al. (in press) also concluded that their mismatch effect stemmed from abstract priming resulting from specific characteristics of the primes rather than from the activation of exemplars. Together, these findings show that we should be careful in concluding that effects are caused by the activation of exemplars, given that general facilitation due to the encoding circumstances may produce similar outcomes. Chapter 2 calls for replication of previous studies with stimuli from different talkers.

How to interpret the reversed amplitude pattern in the EEG signal in Chapter 3 was less obvious, because ERP components are associated with different functions. Thus, unlike speeded or more accurate behavioral responses, a certain amplitude pattern cannot easily be linked to a facilitation in processing. One possibility is that, like in Chapter 2, these ERP data reflect abstract priming. Future research should further investigate the origins of mismatch effects in both behavior and EEG.

Experimental measures

Exemplar effects typically have small effect sizes (e.g., in participants' reaction times, the effects' sizes often range in tens of milliseconds). It may therefore be that in some of the various experiments that produced null results in the literature, exemplar effects were present, but were too weak to be detected in participants' reaction times and/or response accuracy.

This may also have been the case for the behavioral data of Chapter 3. There is a discrepancy between the behavioral findings for the old-new judgment task in Chapters 2 and 3: exemplar effects arose in Chapter 2, while they did not surface in Chapter 3. Exemplar effects were probably present in the behavioral data of the old-new judgment task in Chapters 2 and 3 alike, but could only be detected in Chapter 2. The most important reason to believe the exemplar effects were present in the behavioral data for the old-new judgment task in Chapter 3 as well is that the effects surfaced in the additional experimental measure used alongside behavioral measures

(EEG). A second reason is that patterns in the response accuracy data of Chapter 3, although not reaching statistical significance, trended in the expected direction (i.e., numerically more accurate responses in the match condition). A third reason is that the number of behavioral observations was considerably lower in Chapter 3 than in Chapter 2. Chapter 2 tested 93 participants on 36 experimental words for the old-new judgment task (yielding 3348 datapoints), while in Chapter 3, the final dataset for the old-new judgment task contained data from 22 participants tested on 64 experimental words (yielding 1408 datapoints). The reason the sizes of the datasets in the two studies differed to such a large extent is twofold. First, Chapter 2 tested more participants because of the additional manipulation with three different prime tasks in that study (i.e., there were around 30 participants per prime task condition). Second, in Chapter 3, quite a large number of participants (11 for the old-new judgment task) had to be excluded because of an excessive number of artefacts in their EEG data, and because of technical issues. Assuming similar variance in the behavioral data of the two studies, Chapter 2 could therefore detect smaller exemplar effects elicited by an old-new judgment task than Chapter 3 could. The two studies thus show that exemplar effects for an old-new judment task have small effect sizes, and that sufficient statistical power is required to detect them. This is likely true for exemplar effects elicited by other tasks as well. Future studies should therefore conduct experiments with greater statistical power than the experiments in Chapter 2 had. Additionally, when EEG is used as experimental measure, a larger buffer for participant exclusion is recommended (leaving around 30 rather than 20 participants per task group in the final analysis).

Another example of an effect that may have been present, but was too small to be detected was the null result for prime tasks in Chapter 2. This null finding stands in contrast with positive findings reported by Goldinger (1996) and Theodore et al. (2015), who observed larger exemplar effects for perceptual tasks. To convincingly show that the occurrence of exemplar effect relates to characteristics of episodic memory, the findings by Goldinger and Theodore et al. should therefore be replicated in future research. As before (in Chapter 3), this null result may be due to a lack of statistical power. For instance, Goldinger tested 35 participants on 150 stimulus words, whereas Chapter 2 tested 31 participants on no more than 36 stimulus words per prime task. Assuming similar variance in the data of the two studies, Goldinger could therefore detect smaller effects than Chapter 2 could. Greater statistical power in a future experiment could be achieved by testing more participants, and to a lesser extent more experimental items, as the lag between primes and targets should remain small (as was shown by Hanique et al., 2013), which is not possible with a large number of experimental items. For instance, an experiment with 50 participants on 40 stimulus words per prime task may be able to reveal differences between various prime tasks.

In Chapter 3, exemplar effects were investigated with a method which may more adequately capture them: EEG (electroencephalography). EEG measures cognitive processes more directly than behavioral methods do, because they do not reflect an accumulation of processes, but tap into a single processing stage with millisecond-by-millisecond resolution. Alongside EEG, behavioral measures were collected, which allowed for a comparison of methods. A remarkable difference in detection power between the measures appeared: the EEG data showed clear exemplar effects in two of the three analysis windows, while the behavioral data showed no exemplar effects at all. This suggests that EEG is more sensitive to measure exemplar effects than behavioral measures are. Moreover, the discrepancy between the behavioral and ERP findings hints that exemplar effects may have been present in previous behavioral experiments that produced null results, but that they were to weak to emerge in participants' overt behaviors. Based on Chapter 3, future research is advised to test for exemplar effects with more sensitive methods such as EEG. For comparability with previous behavioral research, behavioral measures should be collected simultaneously (while using sufficient statistical power). In addition, the simultaneous collection of behavioral measures in neuroimaging studies may provide complementary information, and allows researchers to investigate if their neural effects correlate with behavioral effects.

The conclusion that exemplars are represented in episodic memory because exemplar effects mostly arise when participants use their episodic memories is somewhat indirect. More compelling evidence for this claim would be significant activation of the brain areas associated with episodic memory (e.g., the hippocampus) when listeners' behavior shows reliance on exemplars. To investigate this in more detail, an fMRI study using a noise manipulation (like in Chapter 4) in combination with the repetition suppression paradigm (whereby repeated exposure to the same stimulus results in an attenuated brain response in cortical regions that are activated during the processing of that stimulus, e.g., Grill-Spector et al., 1999) was conducted. This study is currently in its analysis stage; results so far suggest small exemplar effects in noise in brain areas related to episodic memory. The next steps are to target a more specific analysis to the regions of interest (ROIs), and to correlate the neural exemplar effects with participants' exemplar effects in their overt behavior. From a methodological point of view, this study shows that neuroimaging methods such as fMRI are a promising tool to answer our research questions in a more direct way, focusing on the activity in specific brain regions.

Conclusions

In support of the hypothesis that exemplars are based in episodic memory, the literature review and the experiments reported in this dissertation show that exemplar effects are more likely to arise when participants are instructed (as in Chapters 2 and 3) or encouraged (as in Chapters 4 and 5) to use their episodic memories. Moreover, the experiments reported in Chapter 5 show that exemplar effects may arise for variation types that are unfamiliar because they do not appear in the listeners' L1.

The studies in this dissertation thus offer different pieces of evidence to suggest that exemplars are based in episodic memory rather than in the mental lexicon. As such, the studies contribute significantly to our understanding of the nature of exemplars. The lack of exemplar effects as well as the reversed effects that arose in some of the experiments suggest that speech comprehension not only involves exemplars, but also abstract representations. This dissertation therefore supports a hybrid model with exemplars based in episodic memory, and abstract representations in the mental lexicon. Importantly, the representation of exemplars in episodic memory implies that speech comprehension involves multiple memory systems in parallel.

The representation of exemplars in episodic memory as is evident from this dissertation suggests a smaller role for exemplars than has been assumed previously. Everyday listening circumstances likely involve little reliance on episodic memory, as listeners process speech for meaning. The fact that exemplar effects only in part generalized to more natural listening circumstances than those typically tested in psycholinguistic experiments further shows that the role for exemplars in everyday speech comprehension is likely to be small. The evidence presented in this dissertation on the episodic nature of exemplars thus has far-reaching consequences for our view of the memory systems involved in speech comprehension.

Nederlandse samenvatting

Een bekende vraag in onderzoek naar taalverwerking is hoe de uitspraak van woorden opgeslagen zit in het hoofd van luisteraars. Deze dissertatie buigt zich ook over die vraag. Vanouds nemen theorieën over taalverwerking aan dat de uitspraak van een woord is opgeslagen als een enkele abstracte representatie. Zo'n abstracte representatie bestaat uit een reeks abstract symbolen, zoals bijvoorbeeld fonemen. Van belang is dat een abstracte representatie geen informatie bevat over wie de spreker was, over hoe de spreker het woord precies uitsprak, en in wat voor situatie hij of zich bevond.

Recenter zijn er ook theorieën ontwikkeld die aannemen dat de uitspraak van woorden in het hoofd van de luisteraar zit opgeslagen als vele voorbeelden, zogeheten 'exemplars'. Deze vele exemplars van een woord clusteren samen als een wolk voor dat woord, waarin woorden die sterk op elkaar lijken zich dicht bij elkaar bevinden. Exemplars zijn fonetisch gedetailleerd, en bevatten wél informatie over door wie en hoe een woord precies uitgesproken werd. Tussen deze twee extremen bevinden zich de zogeheten hybride theorieën, waarin abstracte representaties en exemplars gecombineerd worden. De aanname van dit soort theorieën is dat luisteraars beide types representaties voor de uitspraak van woorden opgeslagen hebben.

Het idee dat de uitspraak van woorden (ook) opgeslagen is als exemplars wordt voornamelijk ondersteund door bevindingen van auditieve priming experimenten met woordrepetitie. In dit soort experimenten luisteren proefpersonen naar gesproken woorden die twee keer voorkomen. Typisch gezien zijn proefpersonen sneller en/of beter in het herkennen van woorden wanneer ze voor de tweede keer voorkomen (een priming effect). Een groot aantal studies heeft gevonden dat dit voordeel nóg groter is als de twee herhalingen van het woord sterk op elkaar lijken, bijvoorbeeld omdat ze worden uitsproken door dezelfde spreker, of met hetzelfde spreektempo. Er zijn echter ook een heel aantal experimenten bekend waarin zulke 'exemplar effecten' niet optraden. Op dit moment is onduidelijk onder welke omstandigheden exemplar effecten optreden en wanneer niet. Bovendien roepen de nulresultaten in de literatuur de vraag op of exemplars wel zo'n grote rol spelen in spraakverwerking.

Dit proefschrift onderzoekt wat exemplars precies zijn, en wat voor rol ze spelen in spraakverwerking. De centrale hypothese die getest wordt is dat exemplars niet opgeslagen zitten in het mentale lexicon, zoals de meeste theorieën aannemen, maar in plaats daarvan in het episodisch geheugen. Het episodisch geheugen bevat herinneringen aan meegemaakte gebeurtenissen, en is dus niet specifiek voor taal.

Hoofdstuk 2 bekijkt of de exemplar effecten die zijn gerapporteerd in de literatuur gekoppeld kunnen worden aan of proefpersonen hun episodisch geheugen gebruikten of aan de eigenschappen van het episodisch geheugen. Dit lukt in de meeste gevallen, maar tegelijkertijd is het duidelijk dat meer onderzoek nodig is om te bepalen of exemplars in het episodisch geheugen opgeslagen zitten. In ditzelfde hoofdstuk werden daarom twee experimenten uitgevoerd om verder te onderzoeken of exemplars in het episodisch zitten.

In deze experimenten werden woorden herhaald in dezelfde of in een andere stem, en er werden verschillende taken gebruikt. Wanneer proefpersonen woorden voor de eerste keer hoorden (als ‘primes’), deden ze taken waarbij ze op de uitspraak of op de betekenis van woorden moesten letten. Vanuit theorieën over het episodisch geheugen is bekend dat herinneringen langer in het episodisch geheugen blijven als proefpersonen op aspecten van woorden zoals de uitspraak in plaats van de betekenis letten. Wanneer het woord voor de tweede keer voorkwam (als ‘target’) moesten proefpersonen aangeven of ze dit woord eerder in het experiment hadden gehoord (een ‘old-new’ taak), of aangeven of de woorden op levende of niet-levende zaken sloegen (een semantische taak). Alleen de old-new taak vereist het raadplegen van episodisch geheugen. Als exemplars in het episodisch geheugen zitten, verwachten we daarom de duidelijkste exemplar effecten wanneer proefpersonen de primes op basis van de uitspraak classificeerden, en de targets op basis van of ze eerder voorkwamen. De resultaten lieten zien dat exemplar effecten duidelijk optraden voor de old-new taak op de targets, terwijl de effecten er niet waren voor de semantische taak. De effecten hingen niet af van welke taak proefpersonen deden tijdens het luisteren naar de de primes, waarschijnlijk door een gebrek aan statistische power. Samen zijn de resultaten van voorgaande literatuur en de experimenten in dit hoofdstuk in overeenstemming met de hypothese dat exemplars in het episodisch geheugen opgeslagen zitten.

Ook Hoofdstuk 3 onderzoekt of exemplar effecten duidelijker optreden in een old-new taak (waarvoor het raadplegen van het episodisch geheugen nodig is) dan in een semantische taak (waarvoor het episodisch geheugen niet geraadpleegd hoeft te worden). Naast het meten van hoe snel en accuraat mensen woorden kunnen herkennen (gedragsmaten) werd in dit hoofdstuk een methode gebruikt die mogelijk gevoeliger is voor het meten van exemplar effecten: EEG (elektro-encefalografie). Waar voor het meten van reactietijden en accuratesse een openlijke respons van proefpersonen nodig is die het resultaat is van een optelsom van cognitieve processen, wordt EEG als een directere meetmethode beschouwd omdat deze methode een enkel cognitief proces in kaart kan brengen. In de EEG data traden duidelijke exemplar effecten voor de old-new taak op (een replicatie van de gedragsdata gerapporteerd in Hoofdstuk 2), terwijl de resultaten voor de semantische taak minder éénduidig te interpreteren waren. Van belang is dat de effecten alleen meetbaar waren in de EEG data, terwijl ze in de gedragsdata niet te detecteren waren. Deze laatste bevinding suggereert dat EEG een geschiktere methode is om exemplar effecten mee te onderzoeken dan dat gedragsmaten zijn. Ook kan dit verschil in gevoeligheid tussen de twee methoden een deel van de nulresultaten in de literatuur verklaren waar vooral gebruik werd gemaakt van gedragsmaten.

In Hoofdstuk 4 worden luisteromstandigheden gemanipuleerd. In priming experimenten met moeilijke luisteromstandigheden kan het gebruik van episodisch geheugen helpen bij woordherkenning. Als zodanig moedigen moeilijke luisteromstandigheden proefpersonen aan om hun episodisch geheugen te raadplegen. Als exemplars in het episodisch geheugen opgeslagen zitten, zouden exemplar effecten het meest duidelijk moeten optreden onder moeilijke luisteromstandigheden. Hoofdstuk 4 maakt enerzijds gebruik van duidelijke spraak en anderzijds van spraak in achtergrondruis. De resultaten lieten zien dat exemplar effecten alleen optraden bij achtergrondruis. Dit waren precies de omstandigheden waarin proefpersonen werden aangemoedigd om hun episodisch geheugen te raadplegen. Dit hoofdstuk is daarom ook in overeenstemming met de hypothese dat exemplars in het episodisch geheugen opgeslagen zitten.

Hoofdstuk 5 maakt gebruik van luisteraarpopulaties die verschillen in hun luisterbelasting. Twee lexicale decisie-experimenten werden uitgevoerd met moedertaalluisteraars (Engelsen) en twee groepen niet-moedertaalluisteraars (Spanjaarden en Nederlanders). Het is bekend dat luisteren in een vreemde taal meer moeite kost dan in de moedertaal, en het was onduidelijk of exemplar effecten ook onder die luisteromstandigheden optreden. Dat wordt onderzocht in Hoofdstuk 5. Dit hoofdstuk belicht ook de vraag wat voor soort informatie er opgeslagen zit in exemplars, en wat dat ons kan vertellen over waar exemplars opgeslagen zitten. In het episodisch geheugen kan alle soorten informatie opgeslagen worden, terwijl het mentale lexicon waarschijnlijk beperkingen heeft gerelateerd aan de moedertaal van luisteraars op de informatie die opgeslagen kan worden.

Er werd getest voor exemplar effecten van sprekerstem en van reductie (het verkorten of weglaten van klanken in lopende spraak). Variatie in sprekerstem is niet taalspecifiek, en was dus bekend voor alle luisteraars. Hoe woorden gereduceerd worden is voor een groot deel wél taalspecifiek. Eén van de groepen niet-moedertaalluistaars (Nederlanders) was bekend met het soort reducties in het Engels dat getest werd, terwijl de andere groep (Spanjaarden) dat helemaal niet was. Een belangrijke bevinding van Hoofdstuk 5 is dat exemplar effecten optraden in het experiment met reductie voor de Spanjaarden. Dit laat zien dat exemplars ook informatie kunnen bevatten die nauwelijks bekend is vanuit de moedertaal van een luisteraar. Dit resultaat wijst naar het episodisch geheugen als opslagplaats voor exemplars. Ook bleek dat het optreden van exemplar effecten afhangt van hoeveel cognitieve middelen luisteraars ter beschikking hebben (waarbij exemplar effecten eerder optreden als er veel middelen beschikbaar zijn) en van hoezeer de variatie in een experiment luisteraars opvalt (waarbij het optreden van exemplar effecten waarschijnlijker is naarmate de variatie meer opvalt).

Samen bieden de hoofdstukken in dit proefschrift verschillende types evidentie die erop wijzen dat exemplars in het episodisch geheugen opgeslagen zitten in plaats van in het mentale lexicon. De inventarisatie van de literatuur liet zien dat het optreden van exemplar effecten kan worden gerelateerd aan eigenschappen van het episodisch geheugen en aan of luisteraars hun episodisch geheugen raadplegen. De experimenten lieten ook zien dat exemplar effecten het meest duidelijk optreden wanneer proefpersonen hun episodisch geheugen raadplegen, omdat de taak ze dat opdraagt, of omdat het ze helpt bij het herkennen van woorden die lastig te verstaan zijn. De studie met niet-moedertaalluisteraars liet bovendien zien dat informatie die nauwelijks bekend is uit de moedertaal ook opgeslagen kan worden in exemplars.

In de studies in deze dissertatie waren er ook gevallen waarin luisteraars voor spraakverwerking vooral gebruik leken te maken van abstracte representaties. De bevindingen van dit proefschrift steunen daarom een theoretisch model waarin exemplars, opgeslagen in het episodisch geheugen, gecombineerd zijn met abstracte representaties die in het mentale lexicon opgeslagen zitten. Dit proefschrift draagt bij aan onze kennis van wat exemplars precies zijn, en onder welke omstandigheden ze een rol spelen. De bevinding dat exemplars in het episodisch geheugen opgeslagen zitten impliceert dat er, anders dan eerder aangenomen werd, meerdere geheugensystemen tegelijk bij spraakbegrip betrokken zijn. Aangezien het episodisch geheugen alleen wordt geraadpleegd onder bepaalde omstandigheden, suggereren de bevindingen in dit proefschrift ook dat de rol van exemplars in spraakverwerking minder groot is dan dat eerder gedacht werd.

References

Baayen, R. H. & Milin, P. (2010). Analyzing Reaction Times. *International Journal of Psychological Research*, *3*(2), 12–28. doi:10.21500/20112084.807

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX Lexical Database. Release 2 (CD-ROM)*. Philadelphia, Pennsylvania: Linguistic Data Consortium, University of Pennsylvania.

Bates, D. M., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi:10.18637/jss.v067.i01

Boersma, P. & Weenink, D. (2018). Praat: doing phonetics by computer (Version 6.0.43).

Booth, J. R., Burman, D. D., Meyer, J. R., Gitelman, D. R., Parrish, T. B., & Mesulam, M. M. (2002). Functional anatomy of intra- and cross-modal lexical tasks. *NeuroImage*, *16*(1), 7–22. doi:10.1006/nimg.2002.1081

Borghini, G. & Hazan, V. (2018). Listening effort during sentence processing is increased for non-native listeners: A pupillometry study. *Frontiers in Neuroscience*, *12*, 1–13. doi:10.3389/fnins.2018.00152

Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception and Psychophysics*, *61*(2), 206–219. doi:10.3758/BF03206883

Bradlow, A. R. & Pisoni, D. B. (1999). Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors. *The Journal of the Acoustical Society of America*, *106*(4), 2074–2085. doi:10.1121/1.427952

Brand, S. & Ernestus, M. (2018). Listeners' processing of a given reduced word pronunciation variant directly reflects their exposure to this variant: Evidence from native listeners and learners of French. *Quarterly Journal of Experimental Psychology*, *71*(5), 1240–1259. doi:10.1080/17470218.2017.1313282

Bürki, A. & Gaskell, M. G. (2012). Lexical representation of schwa words: Two mackerels, but only one salami. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(3), 617–631. doi:10.1037/a0026167

Bybee, J. L. (2001). *Phonology and language use*. Cambridge: Cambridge University Press. doi:https://doi.org/10.1017/CBO9780511612886

Campeanu, S., Craik, F. I. M., Backer, K. C., & Alain, C. (2014). Voice reinstatement modulates neural indices of continuous word recognition. *Neuropsychologia*, *62*, 233–244. doi:10.1016/j.neuropsychologia.2014.07.022

Capa, R. L., Audiffren, M., & Ragot, S. (2008). The interactive effect of achievement motivation and task difficulty on mental effort. *International Journal of Psychophysiology*, *70*, 144–150. doi:10.1016/j.ijpsycho.2008.06.007

Church, B. A. & Schacter, D. L. (1994). Perceptual specificity of auditory priming: implicit memory for voice intonation and fundamental frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(3), 521–533. doi:10.1037/0278-7393.20.3.521

Cook, S. V., Pandža, N. B., Lancaster, A. K., & Gor, K. (2016). Fuzzy nonnative phonolexical representations lead to fuzzy form-to-meaning mappings. *Frontiers in Psychology*, *7*(1345), 1–17. doi:10.3389/fpsyg.2016.01345

Cooper, A. W. & Bradlow, A. R. (2017). Talker and background noise specificity in spoken word recognition memory. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *8*(1), 1–15. doi:http://doi.org/10.5334/labphon.99

Cooper, A. W., Brouwer, S., & Bradlow, A. R. (2015). Interdependent processing and encoding of speech and concurrent background noise. *Attention, Perception, and Psychophysics*, *77*(4), 1342–1357. doi:10.3758/s13414-015-0855-z

Craik, F. I. M. & Kirsner, K. (1974). The effect of speaker's voice on word recognition. *Quarterly Journal of Experimental Psychology*, *26*(2), 274–284. doi:10.1080/14640747408400413

Craik, F. I. M. & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684. doi:10.1016/S0022-5371(72)80001-X

Craik, F. I. M. & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 268–294.

Cutler, A., Eisner, F., McQueen, J. M., & Norris, D. (2010). How abstract phonemic categories are necessary for coping with speaker-related variation. In C. Fougeron, B. Kühnert, & M. D'Imperio (Eds.), *Laboratory phonology 10* (pp. 91–111). Berlin: De Gruyter Mouton. doi:10.1063/1.3033202

Cutler, A. & Robinson, T. (1992). Response time as a metric for comparison of speech recognition by humans and machines. In *Proceedings of the second international conference on spoken language processing: Vol. 1* (pp. 189–192). Alberta: University of Alberta.

Dalby, J. (1986). *Phonetic structure of fast speech in American English* (Doctoral dissertation, Indiana University).

Desroches, A. S., Newman, R. L., & Joanisse, M. F. (2009). Investigating the time course of spoken word recognition: Electrophysiological evidence for the influences of phonological similarity. *Journal of Cognitive Neuroscience*, *21*(10), 1893–1906. doi:10.1162/jocn.2008.21142

Dijkstra, T. & Heuven, W. J. B. V. (2002). The architecture of the bilingual word recognition system: from identification to decision. *Bilingualism: Language and Cognition*, *5*(3), 175–197. doi:10.1017/S1366728902003012

Drozdova, P., Hout, R. V., & Scharenborg, O. (2019). Talker-familiarity benefit in non-native recognition memory and word identification : The role of listening conditions and proficiency. *Attention, Perception, and Psychophysics*, 1–23. doi:10.3758/s13414-018-01657-5Talker-familiarity

Dudai, Y. (2012). The restless engram: Consolidations never end. *Annual Review of Neuroscience*, *35*(1), 227–247. doi:10.1146/annurev-neuro-062111-150500

Dufour, S., Bolger, D., Massol, S., Holcomb, P. J., & Grainger, J. (2017). On the locus of talker-specificity effects in spoken word recognition: an ERP study with dichotic priming. *Language, Cognition and Neuroscience*, *32*(10), 1273–1289. doi:10.1080/23273798.2017.1335421

Dufour, S. & Nguyen, N. (2014). Access to talker-specific representations is dependent on word frequency. *Journal of Cognitive Psychology*, *26*(3), 256–262. doi:10.1080/20445911.2014.890204

Ernestus, M. (2000). *Voice assimilation and segment reduction in casual Dutch: A corpus-based study of the phonology-phonetics interface* (Doctoral dissertation, Utrecht, Nederland).

Ernestus, M., Baayen, R. H., & Schreuder, R. (2002). The recognition of reduced word forms. *Brain and Language*, *81*(1-3), 162–173. doi:10.1006/brln.2001.2514

Ernestus, M., Dikmans, M., & Giezenaar, G. (2017). Advanced second language learners experience difficulties processing reduced word pronunciation variants. *Dutch Journal of Applied Linguistics*, *6*(1), 1–31. doi:10.1075/dujal.6.1.01ern

Ernestus, M. & Warner, N. (2011). An introduction to reduced pronunciation variants. *Journal of Phonetics*, *39*, 253–260. doi:10.1016/S0095-4470(11)00055-6

Federmeier, K. D. & Kutas, M. (2002). Picture the difference: Electrophysiological investigations of picture processing in the two cerebral hemispheres. *Neuropsychologia*, *40*(7), 730–747. doi:10.1016/S0028-3932(01)00193-2

Fox, J. & Monette, G. (2002). *R and S-Plus companion to applied regression*. Thousand Oaks, CA, USA: Sage Publications, Inc.

Fox, J. & Weisberg, S. (2011). *An R companion to applied regression* (2nd ed.). Thousand Oaks, CA, USA: Sage Publications, Inc.

Friedrich, C. K., Kotz, S. A., Friederici, A. D., & Alter, K. (2004). Pitch modulates lexical identification in spoken word recognition: ERP and behavioral evidence. *Cognitive Brain Research*, *20*(2), 300–308. doi:10.1016/j.cogbrainres.2004.03.007

Friedrich, C. K., Kotz, S. A., Friederici, A. D., & Gunter, T. C. (2004). ERPs reflect lexical identification in word fragment priming. *Journal of Cognitive Neuroscience*, *16*(4), 541–552. doi:10.1162/089892904323057281

Furutsuka, T. (1989). Effects of rapid attention switching on the N1-P2 amplitude of the visual event-related potentials. In *Research & clinical center for child development, annual report* (Vol. 11, pp. 55–64).

Glanzer, M. & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory and Cognition*, *13*(8), 8–20. doi:10.1037/0278-7393.16.1.5

Goh, W. D. (2005). Talker variability and recognition memory: instance-specific and voice-specific effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(1), 40–53. doi:10.1037/0278-7393.31.1.40

Goldinger, S. D. (1996). Words and voices: episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1166–1183. doi:10.1037/0278-7393.22.5.1166

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological review*, *105*(2), 251–79. doi:10.1037/0033-295X.105.2.251

Goldinger, S. D. (2007). A complementary-systems approach to abstract and episodic speech perception. *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS 2007)*, (August), 49–54.

Golob, E. J. & Starr, A. (2004). Serial position effects in auditory event-related potentials during working memory retrieval. *Journal of Cognitive Neuroscience*, *16*(1), 40–52. doi:10.1162/089892904322755548

Grill-Spector, K., Kushnir, T., Edelman, S., Avidan, G., Itzchak, Y., & Malach, R. (1999). Differential processing of objects under various viewing conditions in the human lateral occipital complex. *Neuron*, *24*(1), 187–203. doi:10.1016/S0896-6273(00)80832-6

Grohe, A.-K. & Braun, B. (2013). Implicit learning leads to familiarity effects for intonation but not for voice. In F. Bimbot, C. Cerisara, C. Fougeron, G. Gravier, L. Lamel, F. Pellegrino, & P. Perrier (Eds.), *Interspeech 2013: 14th annual con-*

ference of the international speech communication association (pp. 921–924). Lyon, France: ISCA.

Grossberg, S. (1986). The adaptive self-organization of serial order in behavior: Speech, language, and motor control. In E. Schwab & H. Nusbaum (Eds.), *Pattern recognition by humans and machines, vol. 1: Speech perception* (pp. 187–294). New York, NY: Academic Press.

Grossberg, S. & Myers, C. W. (2000). The resonant dynamics of speech perception: Interword integration and duration-dependent backward effects. *Psychological Review*, *107*(4), 735–767. doi:10.1037/0033-295X.107.4.735

Halle, M. (1985). Speculations about the representation of words in memory. *Phonetic linguistics*, 101–114.

Hanique, I., Aalders, E., & Ernestus, M. (2013). How robust are exemplar effects in word comprehension? *The Mental Lexicon*, *8*(3), 269–294. doi:10.1075/ml.8.3.01han

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, *31*, 373–405. doi:10.1016/j.wocn.2003.09.006

Henrichsen, L. E. (1984). Sandhi-variation: A filter of input for learners of ESL. *Language Learning*, *34*(3), 103–123. doi:10.1111/j.1467-1770.1984.tb00343.x

Hink, R. F., Van Voorhis, S. T., Hillyard, S. A., & Smith, T. S. (1977). The division of attention and the human auditory evoked potential. *Neuropsychologia*, *15*(4-5), 597–605. doi:10.1016/0028-3932(77)90065-3

Hintzman, D. L. (1986). 'Schema abstraction' in a multiple-trace model. *Psychological Review*, *93*(4), 411–428. doi:10.1037/0033-295X.93.4.411

Janse, E. (2008). Spoken-word processing in aphasia: effects of item overlap and item repetition. *Brain and Language*, *105*(3), 185–198. doi:10.1016/j.bandl.2007.10.002

Johnson, K. (1997). *Speech perception without speaker normalization: An exemplar model* (K. Johnson & J. Mullennix, Eds.). San Diego: Academic Press.

Johnson, K. (2004). Massive reduction in conversational American English. In K. Yoneyama & K. Maekawa (Eds.), *Spontaneous speech: Data and analysis. proceedings of the 1st session of the 10th international symposium* (pp. 29–54). Tokyo, Japan: The National Institute for Japanese Language.

Kesteren, M. T. R. V., Rijpkema, M., Ruiter, D. J., Morris, R. G., & Fernandez, G. (2014). Building on prior knowledge: Schema-dependent encoding processes relate to academic performance. *Journal of Cognitive Neuroscience*, *26*(10), 2250–2261. doi:10.1162/jocn

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. doi:10.3758/BRM.42.3.643

Krestar, M. L. & McLennan, C. T. (2013). Examining the effects of variation in emotional tone of voice on spoken word recognition. *The Quarterly Journal of Experimental Psychology*, *66*(9), 1793–1802. doi:10.1080/17470218.2013.766897

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience*, *5*, 831–843. doi:10.1038/nrn1533

Kutas, M. & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (ERP). *Annual Review of Psychology*, *62*, 621–647. doi:10.1146/annurev.psych.093008.131123

Kutas, M. & Van Petten, C. (1988). Event-related brain potentail studies of language. In P. Ackles, J. Jennings, & M. Coles (Eds.), *Advances in psychophysiology* (Vol. 3, pp. 139–187). Greenwich, CT: JAI Press, Inc.

Lawrence, M. A. (2016). *ez: Easy analysis and visualization of factorial experiments. R package version 4.4-0*.

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

Luce, P. A. & Lyons, E. A. (1998). Specificity of memory representations for spoken words. *Memory and Cognition*, *26*(4), 708–715. doi:10.3758/BF03211391

Luck, S. J. (2005). Ten simple rules for designing ERP Experiments. In T. C. Handy (Ed.), *Event-related potentials - a methods handbook* (pp. 17–32). Cambridge, MA: MIT press.

Maris, E. & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. doi:10.1016/j.jneumeth.2007.03.024

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*(1-2), 71–102. doi:10.1016/0010-0277(87)90005-9

Marslen-Wilson, W. D. & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*, 29–63. doi:10.1016/0010-0285(78)90018-X

Mattys, S. L. & Liss, J. M. (2008). On building models of spoken-word recognition: When there is as much to learn from natural "oddities"as artificial normality. *Perception and Psychophysics*, *70*(7), 1235–1242. doi:10.3758/PP.70.7.1235

McClelland, J. L. (2013). Incorporating rapid neocortical learning of new schema-consistent information into complementary learning systems theory. *Journal of Experimental Psychology: General*, *142*(4), 1190–1210. doi:10.1037/a0033812

McClelland, J. L. & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86. doi:10.1016/0010-0285(86)90015-0

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457. doi:10.1037/0033-295X.102.3.419

McLennan, C. T. & González, J. (2012). Examining talker effects in the perception of native- and foreign-accented speech. *Attention, Perception, and Psychophysics*, *74*(5), 824–30. doi:10.3758/s13414-012-0315-y

McLennan, C. T. & Luce, P. A. (2005). Examining the time course of indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 306–321. doi:10.1037/0278-7393.31.2.306

McLennan, C. T., Luce, P. A., & Charles-Luce, J. (2003). Representation of lexical form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(4), 539–553. doi:10.1037/0278-7393.29.4.539

McQueen, J. M. (2005). Speech perception. In K. Lamberts & R. Goldstone (Eds.), *Handbook of cognition* (pp. 255–275). London: Sage Publications, Inc.

Mitterer, H. & Tuinman, A. (2012). The role of native-language knowledge in the perception of casual speech in a second language. *Frontiers in Psychology*, *3*, 249. doi:10.3389/fpsyg.2012.00249

Morano, L., Ten Bosch, L., & Ernestus, M. (in press). Looking for exemplar effects: testing the comprehension and memory representations of r'duced words in Dutch learners of French. In S. Fuchs, A. Rochet-Capella, & J. Cleland (Eds.), *Speech perception and production: Learning and memory* (pp. 1–29). Bern.

Näätänen, R. & Piction, T. (1987). The N1 wave of the human electric and magnetic response to sound: A review and analysis of the component structure. *Psychophysiology*, *24*(4), 375–425. doi:10.1111/j.1469-8986.1987.tb00311.x

Naveh-Benjamin, M. & Craik, F. I. M. (1995). Memory for context and its use in item memory: Comparisons of younger and older persons. *Psychology and Aging*, *10*(2), 284–293. doi:10.1037/0882-7974.10.2.284

Norris, D. (1994). Shortlist: a connectionist model of continuous speech recognition. *Cognition*, *52*, 189–234.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57. doi:10.1037/0096-3445.115.1.39

Nygaard, L. C., Burt, S. A., & Queen, J. S. (2000). Surface form typicality and asymmetric transfer in episodic memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1228–1244. doi:10.1037/0278-7393.26.5.1228

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, *2011*. doi:10.1155/2011/156869

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(2), 309–328. doi:10.1037/0278-7393.19.2.309

Papesh, M. H., Goldinger, S. D., & Hout, M. C. (2016). Eye movements reveal fast, voice-specific priming. *Journal of Experimental Psychology: General*, *145*(3), 314–37. doi:10.1037/xge0000135

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. L. Bybee & P. Hopper (Eds.), *Typological studies in language, vol. 45: Frequency and the emergence of linguistic structure* (pp. 137–157). Amsterdam: John Benjamins Publishing Company. doi:10.1075/tsl.45.08pie

Pierrehumbert, J. B. (2002). Word-specific phonetics. In C. Gussenhoven & N. Warner (Eds.), *Papers in laboratory phonology 7* (pp. 101–139). Berlin: Mouton de Gruyter. doi:10.1515/9783110197105.1.101

Pilotti, M., Bergman, E. T., Gallo, D. A., Sommers, M. S., & Roediger-III, H. L. (2000). Direct comparison of auditory implicit memory tests. *Psychonomic Bulletin & Review*, *7*(2), 347–353. doi:10.3758/BF03212992

Pinheiro, J. C. & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. Statistics and computing. New York, NY: Springer. doi:10.1007/b98882

Pufahl, A. & Samuel, A. G. (2014). How lexical is the lexicon? Evidence for integrated auditory memory representations. *Cognitive Psychology*, *70*, 1–30. doi:10.1016/j.cogpsych.2014.01.001

Rabbitt, P. M. A. (1968). Channel-capacity, intelligibility and immediate memory. *Quarterly Journal of Experimental Psychology*, *20*(3), 241–248.

Ramus, F., Peperkamp, S., Christophe, A., Jacquemot, C., Kouider, S., & Dupoux, E. (2010). A psycholinguistic perspective on the acquisition of phonology. In C.

Fougeron, B. Kühnert, M. D'Imperio, & N. Vallée (Eds.), *Laboratory phonology 10: Variation, phonetic detail and phonological representation* (pp. 311–340). Berlin: De Gruyter Mouton.

RCoreTeam. (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Saldaña, H. M., Nygaard, L. C., & Pisoni, D. B. (1996). Encoding of visual speaker attributes and recognition memory for spoken words. In D. Stork & M. Hennecke (Eds.), *Speechreading by humans and machines. nato asi series (series f: Computer and systems science), vol. 150* (pp. 275–281). Berlin, Heidelberg: Springer. doi:10.1007/978-3-662-13015-5_21

Schacter, D. L. & Church, B. A. (1992). Auditory priming: Implicit and explicit memory for words and voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(5), 915–930. doi:10.1037/0278-7393.18.5.915

Scheffert, S. M. (1998). Contributions of surface and conceptual information to recognition memory. *Perception and Psychophysics*, *60*(7), 1141–1152. doi:10.3758/BF03206164

Schild, U., Becker, A. B., & Friedrich, C. K. (2014a). Phoneme-free prosodic representations are involved in pre-lexical and lexical neurobiological mechanisms underlying spoken word processing. *Brain and Language*, *136*, 31–43. doi:10.1016/j.bandl.2014.07.006

Schild, U., Becker, A. B., & Friedrich, C. K. (2014b). Processing of syllable stress is functionally different from phoneme processing and does not profit from literacy acquisition. *Frontiers in Psychology*, *5*, 1–12. doi:10.3389/fpsyg.2014.00530

Shockey, L. (2003). *Sound patterns of spoken English*. Oxford: Blackwell. doi:10.1002/9780470758397

Sommers, M. S. & Barcroft, J. (2006). Stimulus variability and the phonetic relevance hypothesis: Effects of variability in speaking style, fundamental frequency, and speaking rate on spoken word identification. *The Journal of the Acoustical Society of America*, *119*(4), 2406–2416. doi:10.1121/1.2171836

Squire, L. R. & Alvarez, P. (1995). Retrograde amnesia and memory consolidation: a neurobiological perspective. *Current Opinion in Neurobiology*, *5*(2), 169–177. doi:10.1016/0959-4388(95)80023-9

Strori, D., Zaar, J., Cooke, M., & Mattys, S. L. (2018). Sound specificity effects in spoken word recognition: The effect of integrality between words and sounds. *Attention, Perception, and Psychophysics*, *80*(1), 222–241. doi:10.3758/s13414-017-1425-3

Sumner, M. & Samuel, A. G. (2005). Perception and representation of regular variation: The case of final /t/. *Journal of Memory and Language*, *52*(3), 322–338. doi:10.1016/j.jml.2004.11.004

Theodore, R. M., Blumstein, S. E., & Luthra, S. (2015). Attention modulates specificity effects in spoken word recognition: Challenges to the time-course hypothesis. *Attention, Perception, and Psychophysics*, *77*, 1674–1684. doi:10.3758/s13414-015-0854-0

Torreira, F. & Ernestus, M. (2011). Realization of voiceless stops and vowels in conversational French and Spanish. *Laboratory Phonology*, *2*(2), 331–353. doi:10.1515/labphon.2011.012

Trofimovich, P. (2005). Spoken-word processing in native and second languages: An investigation of auditory word priming. *Applied Psycholinguistics*, *26*(4), 479–504. doi:10.1017/S0142716405050265

Trubetzkoy, N. S. (1939). *Grundzüge der Phonologie* (1969 trans). Berkeley: University of California Press.

Tucker, B. V. & Ernestus, M. (2016). Why we need to investigate casual speech to truly understand language production, processing and the mental lexicon. *The Mental Lexicon*, *11*(3), 375–400. doi:10.1075/ml.11.3.03tuc

Tuft, S. E., McLennan, C. T., & Krestar, M. L. (2016). Hearing taboo words can result in early talker effects in word recognition for female listeners. *Quarterly Journal of Experimental Psychology*, *71*(2), 435–448. doi:10.1080/17470218.2016.1253757

VanRullen, R. (2011). Four common conceptual fallacies in mapping the time course of recognition. *Frontiers in Psychology*, *2*(365). doi:10.3389/fpsyg.2011.00365

Werker, J. F. & Logan, J. S. (1985). Cross-language evidence for three factors in speech perception. *Perception and Psychophysics*, *37*(1), 35–44. doi:10.3758/BF03207136

Wilson, E. B. (1927). Probable interference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, *22*(158), 209–212. doi:10.1080/01621459.1927.10502953

Winters, S., Lichtman, K., & Weber, S. (2013). The role of linguistic knowledge in the encoding of words and voices in memory. In E. Voss, S. Tai, & Z. Li (Eds.), *Selected proceedings of the 2011 second language research forum: Converging theory and practice* (pp. 129–138). Sommerville, MA, USA: Cascadilla Proceedings Project.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*(3), 441–517. doi:10.1006/jmla.2002.2864

Appendix

Chapter 2

Table A1: Stimulus words occurring in Experiments 1 and 2 of Chapter 2 with their English translations.

Experimental words		Fillers	
anker	‘anchor’	broeder	‘friar’
atlas	‘atlas’	butler	‘butler’
auto	‘car’	cijfer	‘digit’
banjo	‘banjo’	dame	‘lady’
beugel	‘brace’	drempel	‘threshold’
cello	‘cello’	hertog	‘duke’
emmer	‘bucket’	hommel	‘bumblebee’
filter	‘filter’	kennel	‘kennel’
foto	‘photograph’	kikker	‘frog’
gordel	‘seatbelt’	komma	‘comma’
harnas	‘armour’	koning	‘king’
heuvel	‘hill’	korrel	‘grain’
hoepel	‘hoop’	krekel	‘cricket’
iglo	‘igloo’	lama	‘llama’
kajak	‘kayak’	meester	‘teacher’
keuken	‘kitchen’	mijter	‘mitre’
klepel	‘clapper’	moeder	‘mother’
koffie	‘coffee’	mossel	‘mussel’
lepel	‘spoon’	oester	‘oyster’
luifel	‘shed’	oma	‘grandmother’
masker	‘mask’	otter	‘otter’
pijler	‘pillar’	panda	‘panda’
raster	‘grid’	peuter	‘toddler’
sabel	‘sabre’	pleister	‘bandaid’
schaduw	‘shadow’	poedel	‘poodle’
sikkel	‘sickle’	priester	‘priest’
snorkel	‘snorkel’	puber	‘adolescent’
spatel	‘spatula’	regel	‘rule’
steiger	‘landing stage’	reiger	‘heron’
tafel	‘table’	rubber	‘rubber’
toren	‘tower’	spiegel	‘mirror’
vijzel	‘mortar’	tralie	‘bar’
villa	‘villa’	varken	‘pig’
zolder	‘attic’	veulen	‘foal’
zwabber	‘mop’	vogel	‘bird’
zwachtel	‘bandage’	wapen	‘weapon’

Chapter 3

Table A2: Stimulus words occurring in the experiment reported in Chapter 3 with their English translations.

Experimental words				**Fillers**			
anker	‘anchor’	koffie	‘coffee’	adder	‘adder’	monnik	‘munk’
atlas	‘atlas’	komma	‘comma’	arend	‘eagle’	mossel	‘mussel’
auto	‘car’	korrel	‘grain’	bever	‘beaver’	ober	‘waiter’
banjo	‘banjo’	ladder	‘ladder’	bode	‘messenger’	oester	‘oyster’
beker	‘beaker’	lepel	‘spoon’	broeder	‘friar’	oma	‘grandmother’
beugel	‘brace’	luifel	‘shed’	butler	’butler’	orka	‘orca’
bezem	‘broom’	masker	‘mask’	dame	‘lady’	otter	‘otter’
borstel	‘brush’	mijter	‘mitre’	danser	‘dancer’	paling	‘eel’
buidel	‘pouch’	molen	‘mill’	egel	‘hedgehog’	panda	‘panda’
cello	‘cello’	oven	‘oven’	eland	‘elk’	pater	‘father’
cijfer	‘digit’	pleister	‘bandaid’	ezel	‘donkey’	pelgrim	‘pilgrim’
cirkel	’circle’	poeder	‘powder’	goeroe	‘guru’	peuter	‘toddler’
deksel	‘lid’	raster	‘grid’	hamster	‘hamster’	poedel	‘poodle’
drempel	‘threshold’	regel	‘rule’	heerser	‘ruler’	priester	‘priest’
emmer	‘bucket’	rubber	‘rubber’	hommel	‘bumblebee’	puber	‘adolescent’
filter	‘filter’	sabel	‘sabre’	imker	‘beekeeper’	reiger	‘heron’
foto	‘photograph’	schaduw	‘shadow’	karper	‘carp’	schilder	‘painter’
gieter	‘watering can’	snorkel	‘snorkel’	kater	‘tomcat’	schipper	‘skipper’
gordel	‘seatbelt’	spatel	‘spatula’	kelner	‘waiter’	slager	‘butcher’
halte	‘stop’	spiegel	‘mirror’	kerel	‘guy’	spreker	‘speaker’
harnas	‘armour’	steiger	‘landing stage’	kieviet	‘lapwing’	tante	‘aunt’
heuvel	‘hill’	stencil	‘stencil’	kikker	‘frog’	tijger	‘tiger’
hoepel	‘hoop’	tafel	‘table’	kleuter	‘toddler’	toekan	‘toucan’
iglo	‘igloo’	toren	‘tower’	kokkel	‘cockle’	trainer	‘trainer’
kabel	‘cable’	tralie	‘bar’	koning	‘king’	varken	‘pig’
kachel	‘heater’	vijzel	‘mortar’	kuiken	‘chick’	veulen	‘foal’
kajak	‘kayak’	villa	‘villa’	lama	‘llama’	vlinder	‘butterfly’
kennel	‘kennel’	wafel	‘waffle’	leerling	‘pupil’	vogel	‘bird’
ketting	‘chain’	wapen	‘weapon’	leider	‘leader’	wezel	‘weasel’
keuken	‘kitchen’	zegel	’seal’	meester	‘teacher’	winnaar	‘winner’
klepel	‘clapper’	zolder	‘attic’	merrie	‘mare’	zwager	‘brother in law’
knuppel	‘club’	zwabber	‘mop’	moeder	‘mother’	zwaluw	‘swallow’

Chapter 4

Table A3: Stimuli occurring in the experiments reported in Chapter 4. Real words are given with their English translations (continues on next page).

Repeated stimuli			**Non-repeated stimuli**		
Real		Pseudo	Real		Pseudo
begieten	‘to water’	bedangen	bedanken	‘to thank’	bedelken
begluren	‘to spy’	bedinken	bedaren	‘to calm down’	bedirven
begraven	‘to bury’	begannen	begeren	‘to desire’	bedoeren
begrijpen	‘to understand’	begoeren	beginnen	‘to start’	begennen
begroeten	‘to greet’	begranzen	begroten	‘to estimate’	begrooien
bekladden	‘to besmirch’	begruien	beheksen	’to jinx’	bekliegen
beklimmen	‘to climb’	beklegen	beheren	‘to manage’	bekreipen
bekransen	‘to garland’	bekonnen	bejagen	‘to hunt’	belamen
bekrassen	‘to scratch’	bekrapen	bekeren	‘to convert’	bemonnen
beschaven	‘to civilize’	bekrempen	bekronen	‘to crown’	benoeten
beschermen	‘to protect’	benotten	beleggen	‘to invest’	benuiden
beschrijven	‘to describe’	bepleuten	belonen	‘to reward’	bepelen
besmeren	‘to smear’	beplonten	bemerken	‘to notice’	bepraven
bestelen	‘to rob’	beporken	bereiden	‘to prepare’	beristen
bestoken	‘to harass’	beschakken	besmetten	‘to contaminate’	beschekken
bestraffen	‘to punish’	beschoeten	bestijgen	‘to ascend’	beslatten
bestralen	‘to irridiate’	besmotten	bestraten	‘to pave’	bespraaien
besturen	‘to drive’	bestermen	betreffen	‘to concern’	bestroeien
betasten	‘to touch’	bestraaien	betwisten	‘to dispute’	betreuden
betrappen	‘to catch’	betaffen	bevolken	‘to populate’	bevengen
bevriezen	‘to freeze’	betroeren	bevrijden	‘to liberate’	bewirken
bevruchten	‘to inseminate’	bevichten	beweren	‘to claim’	bezekken
bezingen	‘to chant’	bevrammen	bewerken	‘to manipulate’	bezieten
bezorgen	‘to deliver’	bezeiten	bezetten	‘to occupy’	bezoelen
verbannen	‘to banish’	verbloffen	verbergen	‘to hide’	verbliffen
verbranden	‘to burn	verbrissen	verbouwen	‘to renovate’	verblijmen
vergeven	‘to forgive’	verdechten	verdampen	‘to evaporate’	verbrussen
vergrijzen	‘to age’	verdilgen	verdenken	‘to suspect’	verdetsen
verkiezen	‘to elect’	verdoepen	verduren	‘to endure’	verdirven
verklappen	‘to reveal’	verdrooien	verdwalen	’to get lost’	vergroemen
verkleumen	‘to freeze’	verfalmen	vergoeden	‘to reimburse’	vergussen
verkreuken	‘to crease’	verfrinsen	vergokken	‘to gamble’	verkirsen
vermoeien	‘to tire’	vergippen	vergroeien	‘to coalesce’	verknaren
verprutsen	‘to mess up’	vergoeten	verkleuren	‘to discolor’	verloenken
verslapen	‘to oversleep’	vergreuzen	verknallen	‘to mess up’	verlunken
verslikken	‘to choke’	verguilen	verlangen	‘to desire’	verniemen
verspelen	‘to waste’	verklenen	vermaken	‘to entertain’	verpatten

(Table A3: continued)

Real		Pseudo	Real		Pseudo
versperren	‘to bar’	verknillen	vermengen	‘to mix up’	verpippen
verspreiden	‘to spread’	verkoezen	verplaatsen	‘to move’	verrosten
verstijven	‘to stiffen’	verscharpen	verplichten	‘to force’	versmeuden
verstoten	‘to repudiate’	versnallen	verrekken	‘to strain’	versmieden
vertellen	‘to tell’	verstoemen	verstikken	‘to sufficate’	verspallen
verteren	‘to digest’	verteuven	verstoppen	‘to hide’	vertoelen
vertolken	‘to interpret’	vertiemen	vertakken	‘to branch out’	verwalken
vertragen	‘to delay’	vertilmen	vertalen	‘to translate’	verwijpen
vertrappen	‘to trample’	verwilken	vertoeven	‘to abide’	verzoepen
vertrekken	‘to depart’	verzekken	verzenden	‘to send’	verzwekken
verzachten	‘to soften’	verzwukken	verzinnen	‘to invent’	verzweugen

Chapter 5

Table A4: Stimuli occurring in the experiments reported in Chapter 5. The repeated real words are the experimental words.

Repeated stimuli		Non-repeated stimuli	
Real	Pseudo	Real	Pseudo
balloon	ballee	recall	rekel
banana	benooga	research	resers
belief	beleesh	repair	bedoeren
career	kerame	vanilla	vanole
cassette	kaseet	supply	supplee
cement	semont	deposit	depaset
collapse	coliss	canoe	canee
committee	komanee	result	rezell
debate	debome	selection	selaksin
decline	decloof	salute	saluke
defeat	defoose		
defect	defess		
defence	defots		
degree	degoo		
delay	delow		
design	dezone		
disease	dezoom		
divorce	devees		
domain	domoon		
guitar	guitee		
machine	mechoon		
parade	parogue		
police	poloose		
potato	potono		
safari	saforro		
salami	saleemo		
saloon	saleen		
surprise	suppees		
tobacco	tabodo		
tomato	tomeeno		

Acknowledgments

Completing a PhD project is not a one-woman show. I consider myself lucky to have been surrounded by many supportive people who helped me to make this happen.

First and foremost, I would like to thank my supervisors Mirjam Ernestus and Louis ten Bosch. Mirjam, you taught me about all aspects of academic life. Thank you for challenging me to sharpen my ideas, for the freedom you gave me to make the project my own, for your optimism, and for the many times you gave valuable feedback on my drafts. I wasn't always happy to receive new comments, but they always improved my writing. Louis, my fellow 'Leienaar', thank you for always making time for me, for helping me solve many problems, and for carefully teaching me all I needed to know about data, phonetics, programming and statistics. I appreciate your attempts at making me famous by naming an algorithm after me, and for encouraging me to go on and win Nobel prizes. Especially in the last phase of my PhD project, both of you fully dedicated yourselves to helping me finish (including reading my drafts during many weekend hours), and I greatly appreciate that.

I am also grateful to the members of my manuscript committee, Prof. Roeland van Hout, Dr. Mirjam Broersma, Prof. Aoju Chen, Prof. James McQueen, and Dr. Antje Schweitzer, who took to the time to read and evaluate my thesis, and to provide me with helpful suggestions which allowed me to improve this thesis a lot. I appreciate it.

Huib and Lisa, thank you for being my paranymphs. It will mean a lot to have the two of you by my side during the defense. Huib, you were the best office mate I could have wished for when starting out in the wondrous world of doing a PhD. I enjoyed our many chats, about PhD life and about all sorts of other things. Thank you for the good times, and I am glad we continue to catch up on a regular basis, even though we are often geographically challenged. Lisa, my PhD became a lot more fun when we started hanging out together. Thank you for introducing me to the Nijmegen friend group, for distracting me when needed, for always listening and offering your honest opinion, and for fighting the bice with me. I hope we will continue sharing so many laughs together.

I am grateful to the members of the Speech Comprehension group, first based at the MPI, and later at the Erasmus building. Aurora, Chén, Ellen, Emily, Esther, Huib, Iris, Joe, Juliane, Katherine, Kimberley, Laura, Lisa, Lotte, Malte, Mark, Martijn, Mybeth, Robert, Sascha, Sophie, Thordis, Tim, and Xaver, thank you for your scientific input, and the many fun times in and outside of work! At the Donders, I collaborated with Branka Milivojevic for a project separate from my PhD. I thank her for not only working with me on that project, but also giving me advice on the things related to my PhD.

For my study on native and non-native speech perception in English, I went abroad to test participants from different language backgrounds. Dr. Brechtje Post and Prof. Francis Nolan kindly allowed me to test native English participants in the Phonetics Laboratory at Cambridge University in the UK on two occasions, for which I am grateful. They made feel welcome by inviting me to take part in the group's research and social meetings. I would like to thank lab members Calbert Graham, Elaine Schmidt, and Imme Lammertink for showing me around Cambridge (Calbert and Elaine, remember the time you took me to enjoy the view over the city from a tower, and I was way too scared of heights to even look?). I would also like to thank Prof. Sarah Hawkins for her hospitality when hosting Mirjam and me at her house during my first visit to Cambridge.

To test my Spanish non-native listeners, I was hosted by Prof. José Manuel Pardo at the Escuela Técnica Superior de Ingenieros de Telecomunicación of the Universidad Politécnica de Madrid in Spain. I am grateful that he kindly made his group's sound-proof booth available for my experiments, even though my research is quite far removed from his own. I thank Julian David Echeverry for making me feel at home in the lab by assisting me with all things I could possibly need. I am also very grateful to Manuel Pinilla. At first only being a friend of Jorrit's and hardly knowing me at the time, he didn't hesitate to host me in his home in Madrid during my testing trip, and he took the time to show me the best things to do in the city in my free hours. A particularly nice memory is the time he took me to see a Real Madrid match together with his grandfather in Bernabéu.

During my PhD time, I profited from membership of two graduate schools: IMPRS and GSH. I would like to thank Els, Dirkje, and Kevin for their support offered through IMPRS, and Tanja Döller and Peter through GSH. I enjoyed the educational and social events organized by each of the two graduate schools. I would like to thank Conny, Elliot, Elliott, Franziska, Gabriela, Gwilym, Lotte, Merel, Richard, Sara, and Will for the fun times at IMPRS pizza evenings as well as in and around the MPI. A big thank you also to my fellow 'Nijmeegse Taalmiddag' crew: organizing this event together with you was one of the highlights of my PhD time. At GSH/CLS, I would like to thank Ferdy, Gert-Jan, Ilja, Marten, Remy, Saskia, and Thijs for the many fun moments we shared.

I was able to conduct my studies in a perfectly equipped research environment. I would like to thank the MPI library and the MPI technical group for making this possible, in particular Reiner Dirksmeyer, Alex Dukers, and Johan Weustink. In addition, I am greatly indebted to the little army of student assistants of the Speech Comprehension group available to me, who tested the vast majority of my participants, and who helped me with a variety of other tasks. Also, the studies in this dissertation would not exist without all the people who took part in my experiments. Another thank you therefore goes to all participants tested for this dissertation.

Ashley, Bart, Emma, David, Giulio, Lisa, Maarten, Martine, Nicco, Suzanne, and Rick, thank you for making Nijmegen social life special to me. The many Friday drinks, dinners and all other fun with you guys often sometimes even made me travel to Nijmegen on weekends in addition to my daily weekday commute from Utrecht.

Rosalien, Emma, Nicole, Steven, Leonard, and Willemijn, you guys are one of the reasons Jorrit and I moved back to Leiden. Thank you for distracting me from my thesis during many dinners, drinks and holidays. You knew when *not* to ask about the progress of my PhD, you were patient when I was unavailable to you because I was writing, and you always listened when I needed to blow off steam.

A very big thank you goes to my family. Mama en Ruud, dank voor jullie grenzeloze geloof in mij en mijn kunnen. Papa en Anita, bedankt voor jullie support, begrip en de gesprekken. Olga, Tristan en Evy, wat is het fijn om zulke lieve zussen en broer te hebben. Dank voor alle fijne momenten met jullie. Opa en oma, van kleins af aan zijn jullie misschien wel mijn grootste supporters. Dank hiervoor, ik heb me er erg gesteund door gevoeld. Jacques, Anneke, Jacco, Marieke, Teun, Fenna, Juriaan, Karin en Eva, dank voor jullie grote betrokkenheid en steun. Ook dank voor de weekenden in Zeeland en voor de keren dat jullie me op fiets- en hardlooptochten hebben meegesleurd, het waren welkome afleidingen. Ik voel me heel rijk met jullie allen als familie om me heen.

En natuurlijk Jorrit. Je hebt de wetenschap zelf lang geleden achter je gelaten, en nu ik uitgegroeid ben tot de grootste nerd van ons twee kwam het op mij aan om de eerste doctor in de familie te worden. Dit project heb jij vanaf het begin toegejuicht. Je aarzelde niet om met me van Leiden naar Utrecht te verhuizen zodat ik iedere dag op en neer naar Nijmegen kon, en nam voor lief dat ik vaak precies niet op vakantie kon op momenten dat jij daar tijd voor had vrijgemaakt. Dank hiervoor, en ook voor je gekke grapjes, voor het mij op de hak nemen, voor je coaching tijdens lastige momenten, en bovenal voor je vertrouwen. Ik ben gek op je, en zonder jou was deze dissertatie er nooit gekomen.

Curriculum Vitae

Annika Nijveld was born in 1987 in Voorburg, the Netherlands. She obtained her Bachelor's degree from Leiden University in 2009, after which she completed the Research Master Linguistics at the same university in 2012 (cum laude).

In 2012, she started her PhD project in the Speech Comprehension PI group at the Centre for Language Studies at Radboud University in Nijmegen, funded by NWO. During her PhD time, Annika took several trips to Cambridge (UK) and Madrid (Spain) to test participants for her study involving non-native listeners, and she was involved in the organization of events for both academic and non-academic audiences. Moreover, she supervised a master's thesis in the Research Master Linguistics at Radboud University.

From 2015 to 2017, Annika was appointed at Radboud University to lead a collaborative brain imaging project between the Speech Comprehension PI group and the Memory and Space lab of Christian Doeller at the Donders Centre for Cognitive Neuroimaging. The Memory and Space lab specializes in the neural architecture of human memory. In this project, Annika and her team investigated the neural correlates of exemplar effects in noise using fMRI.

Afterwards, Annika received another appointment at the Centre for Language Studies to finish her PhD project. In 2019, she will start a post-doc position at Phonetics Laboratory of the University of Alberta.

List of Publications

Nijveld, A., Bentum, M., ten Bosch, L., and Ernestus, M. (submitted). The nature and relevance of exemplars in spoken word recognition.

Nijveld, A., ten Bosch, L., and Ernestus, M. (submitted). The use of exemplars differs between native and non-native listening.

Nijveld, A., ten Bosch, and L., and Ernestus, M. (submitted). ERP signal analysis with temporal resolution using a time window bank.

Nijveld, A., Mulder, M., ten Bosch L., and Ernestus, M. (submitted). ERPs but not behavioral measures show that exemplar effects differ depending on whether participants use their episodic memories.

Nijveld, A., ten Bosch, L., Milivojevic, B., Noordenbos, M., Doeller, C. and Ernestus, M. (in prep.). The neural correlates of exemplar effects in speech comprehension in noise: an fMRI study.

Nijveld, A., ten Bosch, L., and Ernestus, M. (2015). Exemplar effects arise in a lexical decision task, but only under adverse listening conditions. *Proceedings of the International Congress of Phonetic Sciences 2015*, Glasgow, Scotland.

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 te 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 Pluymaekers

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

45. Space and iconicity in German sign language.
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 electro-encephalography, 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 place.
    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 Japanse 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 Poellmann

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 Sümer

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

97. Incrementality 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 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, morphosyntatic, 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.
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 Ya’u narrative.
Clair Hill

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

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. Understanding temporal overlap between production and comprehension.
Amie Fairs

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