

Voices in the mental lexicon:

Words carry indexical information that can affect access to their meaning

Efthymia C. Kapnoula

Basque Center on Cognition Brain and Language

&

Arthur G. Samuel

Basque Center on Cognition Brain and Language; Stony Brook University; Ikerbasque

Running Head: TALKER'S VOICE AFFECTS ACCESS TO WORD MEANING

Corresponding Author:

Efthymia C. Kapnoula

Paseo Mikeletegi 69

Basque Center on Cognition, Brain and Language (BCBL)

20009 San Sebastián - Donostia

Spain

kapnoula@gmail.com

## Abstract

The speech signal carries both linguistic and non-linguistic information (e.g., a talker's voice qualities; referred to as *indexical* information). There is evidence that indexical information can affect some aspects of spoken word recognition, but we still do not know whether and how it can affect access to a word's meaning. A few studies support a dual-route model, in which inferences about the talker can guide access to meaning via a route external to the mental lexicon. It remains unclear whether indexical information is *also* encoded within the mental lexicon. The present study tests for indexical effects on spoken word recognition and referent selection *within* the mental lexicon.

In two experiments, we manipulated voice-to-referent co-occurrence, while preventing participants from using indexical information in an explicit way. Participants learned novel words (e.g., *bifa*) and their meanings (e.g., kite), with each talker's voice linked (via systematic co-occurrence) to a specific referent (e.g., *bifa* spoken by speaker 1 referred to a specific picture of a kite). In testing, voice-to-referent mapping either matched that of training (*congruent*), or not (*incongruent*). Participants' looks to the target's referent were used as an index of lexical activation. Listeners looked faster at a target's referent on congruent than incongruent trials. The same pattern of results was observed in a third experiment, when testing was 24 hrs later.

These results show that indexical information can be encoded in lexical representations and affect spoken word recognition and referent selection. Our findings are consistent with episodic and distributed views of the mental lexicon that assume multi-dimensional lexical representations.

Keywords: mental lexicon; indexical effects; word learning; spoken word recognition; visual world paradigm

## Introduction

According to most models of spoken word recognition, lexical representations consist of abstract units, such as phonemes or sets of phonological features (Gaskell & Marslen-Wilson, 1997; McClelland & Elman, 1986; Norris, 1994). However, there is mounting evidence that indexical variability – surface characteristics such as the voice of the speaker – can affect spoken word recognition (Goldinger, 1998; McLennan & Luce, 2005; Pufahl & Samuel, 2014). These findings contradict strictly abstractionist approaches and call for an expanded conceptualization of the mental lexicon that includes both linguistic and indexical information. In fact, there is a growing consensus that a hybrid view of spoken word recognition is needed (Cutler & Weber, 2007; Goldinger, 2007; Sumner, Kim, King, & McGowan, 2014), according to which both abstract representations and episodic memory traces are involved in spoken word recognition. However, a critical open question is whether the episodic traces are stored inside the mental lexicon as part of lexical representations, or if instead they are in a separate memory structure that can be accessed during word recognition.

In the present study, we ask whether indexical information – a talker's voice – can be encoded as part of lexical representations and, as such, directly contribute to spoken word recognition and referent selection. To test this, we employed the visual world paradigm (Allopenna, Magnuson, & Tanenhaus, 1998; Salverda & Tanenhaus, 2017), in which participants see several objects on a computer screen while listening to spoken language (in our case, a target word). We created two contrasting situations in which the visual referent of the target word was congruent with either 1) only phonological or 2) both phonological and indexical information in the speech signal. The difference between the two conditions was the presence or absence of a previously established association between the specific depiction of the target word and the particular speaker's voice. We assessed the time-course of spoken word recognition (using eye-tracking) under these two conditions. Strong abstractionist views would predict no difference between the conditions, because in these models indexical information is filtered out and thus unable to directly affect spoken word recognition. In contrast, if voice information is included in the lexical representation of a referent, there should be a difference in the time-course of lexical activation between the two conditions.

*Abstractionist versus episodic views of the mental lexicon*

According to traditional accounts of spoken word recognition, listeners map speech segments onto abstract word-form representations (see Bowers, 2000; McQueen, Cutler, & Norris, 2006; Samuel, 2011 for discussion). Such abstractionist accounts assume that lexical representations, which are stored in the mental lexicon, are devoid of any extra-phonological information present in the speech signal, such as accent, speech rate, and loudness (commonly referred to as “surface details”). In line with this view, many models of spoken word recognition, including TRACE (McClelland & Elman, 1986), Shortlist (Norris, 1994), and DCM (Gaskell & Marslen-Wilson, 1997, 1999, 2002), assume some form of mapping of the speech input onto abstract features.

In contrast to this view, there is now considerable evidence that indexical variation (such as variation due to differences in talker voice and speaking rate) *can* affect spoken word recognition. Research has established that, rather than filtering out acoustic details of incoming speech, listeners utilize these cues when recognizing words. For example, listeners remember previously heard words better when they are produced by the same speaker (Goldinger, 1996), a speaker of the same gender (Schacter & Church, 1992), or at the same rate (Bradlow, Nygaard, & Pisoni, 1999) as in their previous occurrence. Furthermore, phoneme boundaries can shift as a function of audio or visual cues to speaker sex (Johnson, Strand, & D’Imperio, 1999) or speaker dialect (Niedzielski, 1999; Hay & Drager, 2010).

Many of the aforementioned studies start with an initial encoding phase, during which participants are presented with a series of word stimuli. During the second phase of the experiment, participants are presented with a mix of new and old (i.e., previously presented) words and are asked to indicate whether each stimulus is new or old, and/or whether the voice in which it is spoken is the same or not (Bradlow, Nygaard, & Pisoni, 1999; Cooper & Bradlow, 2017; Palmeri, Goldinger, & Pisoni, 1993). Other studies have used more implicit measures of processing, such as word identification in noise (Goldinger, 1996; González & McLennan, 2007) and lexical decision (Luce & Lyons, 1998). The common finding is that participants perform better when a word is presented in the same voice as in the first presentation (see also Lee & Zhang, 2018). These studies indicate that extra-phonological information is encoded in some form along with phonological information. Interestingly, such information may not be limited to talker-specific information (Goldinger, 1998; Luce, McLennan, & Charles-Luce, 2012;

McLennan & Luce, 2005; Vitevitch & Donoso, 2011), but may also include other aspects of the auditory context in which a word was encountered, such as background noise (Cooper & Bradlow, 2017; Pufahl & Samuel, 2014).

***Do surface details affect referent selection?***

Studies have demonstrated that listeners can use the voice of the speaker to interpret spoken language. For example, Van Berkum, van den Brink, Tesink, Kos, and Hagoort (2008) presented participants with sentences like “Every evening I drink some wine before I go to sleep” spoken by an adult speaker or a young child. In the latter case (in which the message mismatched voice-based inferences about the talker), the critical words (“wine” in this case) elicited an N400 effect, suggesting that listeners used talker-voice information in their semantic interpretation of the sentence.

Additional studies have examined how talker information can be used to guide referent selection. For example, King and Sumner (2015) presented participants with words like “yeast” spoken by either an older African-American male speaker, or a younger White American female speaker and asked participants to report the first word that came to mind. The authors found a significant difference between responses as a function of the identity of the speaker; for example, when spoken by the male speaker, the prompt “yeast” yielded more “bread” responses, whereas when spoken by a female speaker, it yielded more “infection” responses. Similarly, Cai et al. (2017) reported a series of experiments examining how a speaker’s accent can be used to guide disambiguation of semantically ambiguous words. The authors used already-established differential associations between accents and words’ referents (e.g., the word *bonnet* refers to a car part in British English, and to a kind of hat in American English). Consistent with King and Sumner (2015), their results showed that listeners can use talker information (accent in this case) to modulate access to word-meaning.

Similar effects have also been found for newly learned words – a case in which talker-referent association can be experimentally manipulated. Creel and Tumlin (2011) had participants learn similar-sounding novel words (e.g., /beIm/ and /beIn/) as labels for unfamiliar images. Crucially, the two words presented during training were spoken by either the same talker, or two different talkers (i.e., talker A → /beIm/; talker B → /beIn/). At test, participants were able to distinguish the two words more quickly when they were spoken by two different talkers in training than

when they were spoken by the same talker (see also Creel, Aslin, & Tanenhaus, 2008). These results showed that listeners can use voice information to speed up referent selection.

These findings seem to provide support for an episodic view of the lexicon that includes both linguistic information and co-occurring information of various types (Goldinger, 1998).

However, an abstractionist approach could be retained as long as the relevant surface information is stored *outside* the mental lexicon. For example, indexical effects may only apply to early stages of spoken word recognition, limiting their role to the word-form level. According to this interpretation, surface details may be stored within word-form representations, but this information is stripped off before referent selection (see Figure 1.A). That is, one can postulate an intermediate lexical representation in between the word-form and the referent that is abstracted and, in that sense, “clean” from non-phonological information. This idea is consistent with traditional abstractionist accounts (e.g., Norris, 1994), and with more recent accounts, such as that of Cai and colleagues (2017), according to which surface speech details are discarded, but can still be used to bias meaning access in an indirect way. The alternative account is that indexical information is maintained throughout all levels of lexical processing and can, thus, directly affect referent selection (see Figure 1.B).

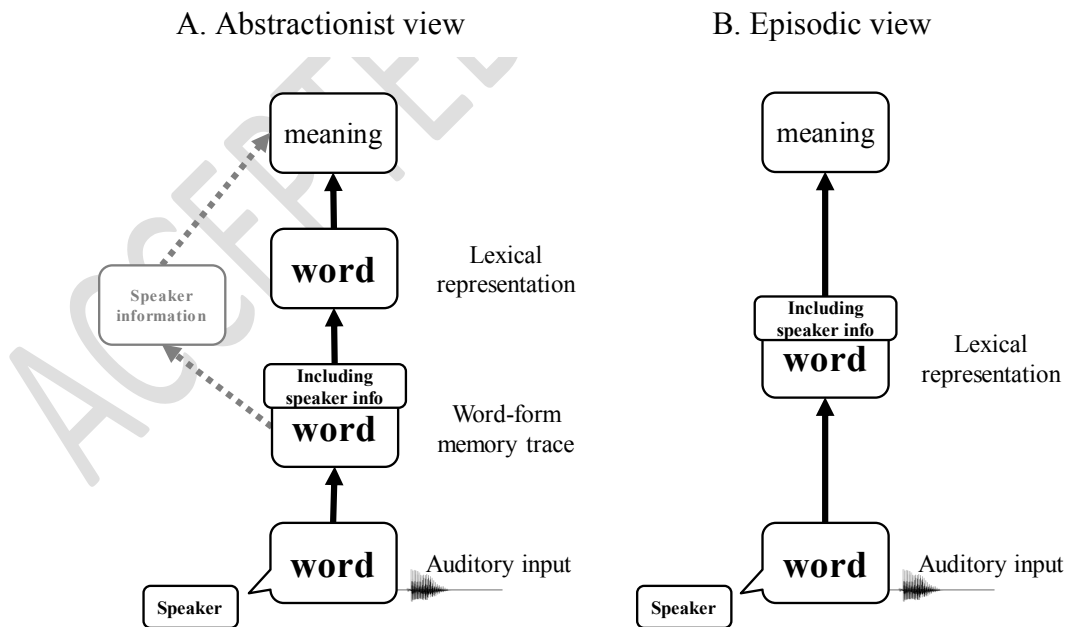


Figure 1. Opposing accounts of indexical effects on referent selection. According to the abstractionist account, auditory input is stripped off of indexical information before being encoded into the mental

lexicon (panel A); according to the episodic account, indexical information is one of the dimensions of lexical representations in the mental lexicon (panel B). Note: Panel A does not depict the claims of any particular abstractionist view, but allows us to visualize a mechanism that is consistent with the abstractionist idea that surface speech details are discarded during spoken word recognition, while also allowing indexical effects to emerge.

Therefore, it is still unknown whether voice information can be encoded as part of lexical representations, and, if so, in what way it may affect how lexical activation and referent selection unfold in real time. The current study was designed to address these questions.

### ***Present study***

The studies reviewed above provide strong evidence that talker information can be used to guide access to meaning. However, they do not speak directly to the issue of whether this information can be encoded as part of a lexical representation because the paradigms have associated different voices with different meanings. In such cases, listeners can use voice information to generate inferences about the talker based on background knowledge (e.g., children don't drink wine, or Americans refer to hats when they talk about bonnets) and, in turn, use those inferences to guide the semantic interpretation of the word. Crucially, the locus of this type of inference-based effect lies *outside* the mental lexicon. Therefore, even if voice information is encoded in the mental lexicon, the effects found in these studies do not provide evidence for or against this.

The goal of the present study was to evaluate whether surface information, such as a talker's voice, can be encoded in the mental lexicon as part of a lexical representation and as such have a direct effect on referent selection. To test this hypothesis, we assessed whether voice information can be implicitly and automatically used, in conjunction with phonological information, in shaping the time course of spoken word recognition.

Our design aimed at creating the experimental conditions in which such an effect, if present, should be most easily observable. For this reason, as in Creel and Tumlin (2011), we used a novel word learning paradigm that allows us to control the kind and amount of indexical information present in the speech signal, as well as the way in which this information is associated with a referent. Critically, and in contrast to previous studies, *voice information was not an informative cue for accessing lexical meaning*. That is, participants were required to map

a given word-form to its meaning regardless of any indexical information or the specific visual instantiation of that meaning – both were irrelevant to the participants' task. For example, if participants had learned that “bifa” refers to kites, they had to map “bifa” to a pictured kite, regardless of what voice said “bifa”, or what particular kite was depicted. At the same time, however, we manipulated the linkage between specific indexical acoustic information (i.e., a speaker's voice) and particular visual instantiations of the word's meaning. As a result, our experimental procedure probes spoken word recognition from the acoustic input all the way to referent selection, while preventing participants from using indexical information in an explicit way. Furthermore, by employing the visual world paradigm, we can monitor the online dynamics of spoken language processing from the earliest parts of the speech signal to the moment when an explicit response is made (in this case, selecting the word's referent). This provides detailed information about how lexical activation unfolds over time starting from its earliest moments.

We conducted two experiments to investigate whether voice information is encoded in lexical representations and as such, can affect referent selection. A third experiment tested whether this effect persists and/or changes over a 24-hr period that includes sleep.

*Overview of experiments.* In all three experiments, Spanish-speaking participants were trained on a set of novel word-forms (e.g., *bifa*), which served as new labels for known objects (e.g., kite [“cometa” in Spanish]). During training, each novel word was spoken in three different voices (e.g., *bifa* spoken by speaker 1, *bifa* spoken by speaker 2, and *bifa* spoken by speaker 3). Similarly, there were three different referents for each novel word form. For example, for *bifa*, the referents were three different images of kites (picture of kite 1, picture of kite 2, picture of kite 3). Unbeknownst to participants, for half of the words they learned, each voice was systematically linked to a specific visual depiction of the word's referent (e.g., *bifa* spoken by speaker 1 always co-occurred with kite picture 1; *indexical condition*). For the other half of the words, no such association was formed (i.e., all three voices co-occurred equally often with all three pictures; *uncorrelated condition*).

At test, participants heard the novel words again and were asked to identify each word's referent. Critically, during testing there was no systematicity between specific talkers' voices and specific depictions of referents. That is, for the items in the indexical condition, the voice-picture co-occurrence sometimes mismatched that of the training (e.g., *bifa* spoken by speaker 1 could now



co-occur with kite picture 2; *incongruent condition*) and sometimes it matched that of training (*congruent condition*). Participants' eye movements to the pictures on the screen were monitored and looks to the target picture were used as a measure of lexical activation of the target word.

Our main hypothesis was tested by comparing word recognition on congruent versus incongruent trials. If listeners rely solely on phonological information to activate words, we should observe no difference between the two conditions. In contrast, if voice information can affect referent selection in a direct way (i.e., via the mental lexicon; see Figure 1B), then activation of the target should be facilitated in the congruent compared to the incongruent condition. Critically, in this design, the lexical meaning is held constant, as the correct response to *bifa* is always to select the picture of a kite, regardless of the particular voice in which it is spoken, or the particular picture of a kite that is present. Thus, participants cannot make explicit use of the indexical information (e.g., via the generation of inferences). In fact, if a participant exclusively relied on voice-to-picture associations, he/she would fail to do the assigned recognition task. If congruent trials nonetheless allow faster access to the referent, that would indicate a direct link between indexical information and lexical meaning.

## **Experiment 1: Can Indexical Information Be Represented in the Lexicon?**

### **Method**

#### ***Participants***

Forty-eight (31 females; mean age = 25.2 years) native speakers of Spanish participated in Experiment 1. Most participants were also fluent in Basque, which was foreseen and taken into account in selecting the stimuli (see *Materials* below). All participants had normal/corrected-to-normal vision and no known hearing or neurological impairments. Participants underwent informed consent and were remunerated for their participation. All experimental procedures were approved by the BCBL ethics committee.

#### ***Design***







Experiment 1 consisted of a *training* and a *testing* phase.














*Training.* Participants learned eight novel words (see Table 1) as labels for eight familiar objects (see Table 2).

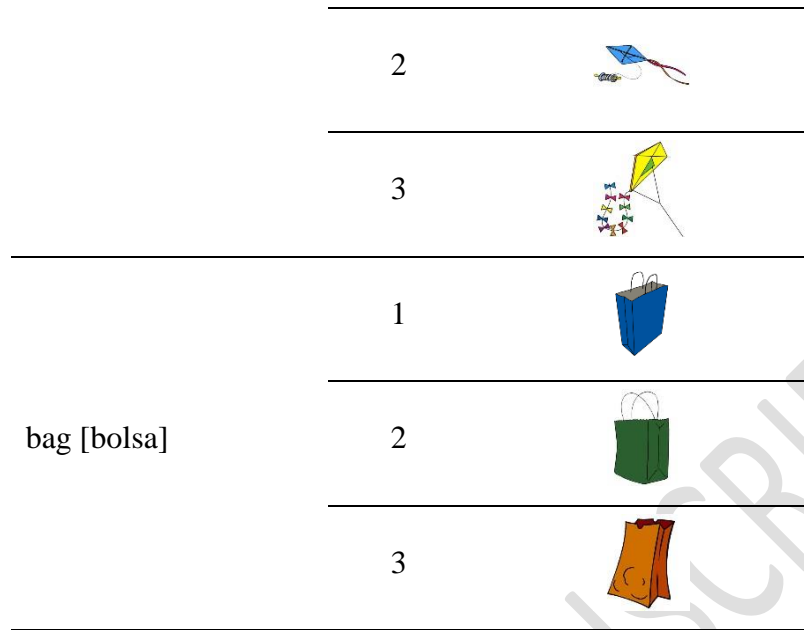
Table 1. *List of novel word stimuli used in Experiments 1, 2, and 3*

| Novel word item | IPA     |
|-----------------|---------|
| bifa            | /bifa/  |
| jito            | /xito/  |
| kena            | /kena/  |
| luda            | /luða/  |
| muko            | /muko/  |
| recho           | /reʧo/  |
| saso            | /saso/  |
| chaga           | /tʃaya/ |

Table 2. *List of objects and corresponding pictures used in Experiments 1, 2, and 3*

| Object name<br>[in Spanish] | Version | Picture  |
|-----------------------------|---------|--|
| dart [dardo]                | 1       |  |
|                             | 2       |  |
|                             | 3       |  |
| table [mesa]                | 1       |  |
|                             | 2       |  |
|                             | 3       |  |

|                     |   |   |
|---------------------|---|---|
|                     | 1 |     |
| comb [peine]        | 2 |     |
|                     | 3 |     |
|                     | 1 |    |
| rope [cuerda]       | 2 |    |
|                     | 3 |    |
|                     | 1 |   |
| bell [campana]      | 2 |   |
|                     | 3 |  |
|                     | 1 |   |
| toaster [tostadora] | 2 |   |
|                     | 3 |   |
| kite [cometa]       | 1 |   |



Each novel word was used as a label of one familiar object (e.g., the novel word “luda” could refer to the object kite). The correspondence between novel words and objects was randomized between participants using a Latin Square. On different trials during the training, each word was presented in one of three different voices and three different images were used to depict each familiar object (e.g., three different pictures of a kite; see Figure 2).

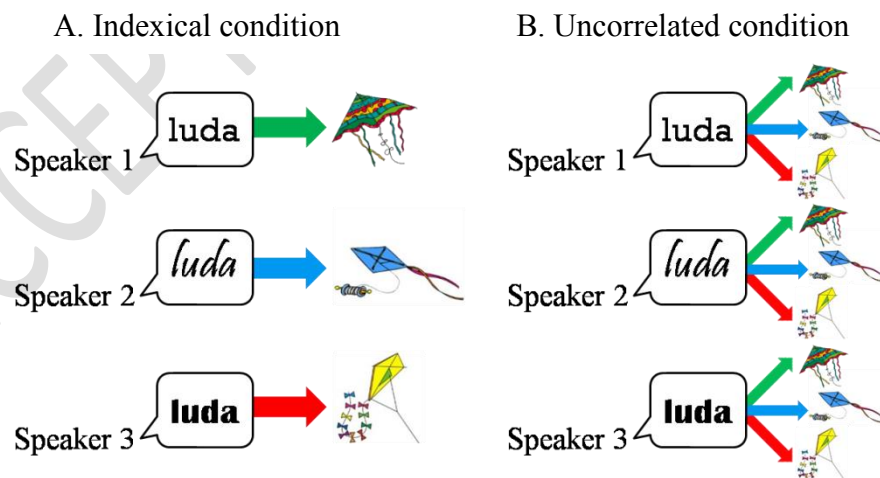


Figure 2. Example of speaker-to-picture correspondences for the indexical (panel A) and the uncorrelated condition (panel B). Note: For a given participant each word (e.g., “luda”) was assigned to one of the two conditions; either the indexical or the uncorrelated.

Crucially, each novel word was assigned to one of two experimental conditions: *indexical* or *uncorrelated* (the assignment of each word to one condition was randomized between participants). Each of the four words in the indexical condition was presented so that whenever it was spoken by a specific speaker it would always correspond to the same picture of its referent. In other words, there was a one-to-one correspondence between voice and picture. For example, if “luda” was the label for the object kite, then “luda” spoken by speaker 1 would always refer to a specific picture of a kite, whereas when spoken by speaker 2, it would always refer to different picture of a kite and so on (see Figure 2A; picture-voice correspondence was counterbalanced between participants). In contrast, no systematic correspondence between voice and picture existed for words in the uncorrelated condition. That is, a word spoken in a given voice had an equal probability of co-occurring with any one of the three pictures of its referent (Figure 2B). During training, each word was presented 63 times (21 times in each voice; i.e., 8 words  $\times$  3 voices  $\times$  21 repetitions = 504 training trials in total), with each of the three pictures of an object used as the corresponding word's referent 1/3 of the time. Importantly, there was no difference between the two experimental conditions in: 1) how many times a word was heard (63), 2) how many times a word was heard in a specific voice (21), or 3) how many times a word referred to a specific picture (21). The only difference was in the systematicity of the voice-to-picture correspondence.

*Testing.* During testing, unlike training, there was no systematic voice-to-picture correspondence for any of the words. A word spoken by a given talker had an equal probability of corresponding to any one of the three pictures of its referent. During testing, each word was heard 27 times (9 times in each voice; i.e., 8 words  $\times$  3 voices  $\times$  9 repetitions = 216 testing trials in total). Each of the three pictures of an object was again used as the corresponding word's referent 1/3 of the time.

### ***Materials***

The novel words were selected from a list of CVCV items. This list was checked by a trained research assistant, who was a Spanish-Basque bilingual. The research assistant identified the items that were nonwords in both Spanish and Basque, but morphologically consistent with Spanish. From the resulting subset, we selected the final stimulus set, making sure that items

were as acoustically dissimilar from each other as possible. Specifically, we made sure that 1) no items shared their first consonant, 2) no items with the same onset manner or place of articulation shared their first vowel, 3) no items shared their second consonant, and 4) no items shared both vowels.

Spoken stimuli were recorded by three native speakers of Spanish (one male and two female) in a sound-attenuated room, sampling at 44,100Hz. The three speakers were selected from sample recordings based on pitch (i.e., fundamental frequency) and overall voice quality with the goal of having three voices that were easily distinguishable from each other (Male speaker pitch<sup>1</sup>: M = 99.6Hz, SD = 15.6; Female 1 speaker pitch: M = 172.7Hz, SD = 6.8; Female 2 speaker pitch: M = 194.4Hz, SD = 8.7).

We collected multiple recordings from each speaker and chose ten recordings per item/per speaker (i.e., 8 items  $\times$  3 speakers  $\times$  10 recordings = 240 recordings) based on sound quality. Chosen recordings were cut, cleaned (background white noise and occasional click/pop sounds removed), and intensity-scaled. Finally, 50 ms of silence were added before and after each word. The average duration of the final stimuli (including the 100 ms of silence) was 662 ms.

Out of these 10 recordings, seven were used in training and the remaining three were used in testing (which audio files were presented in each phase was randomized between participants). This was done for two reasons: 1) using multiple tokens minimizes the possibility of participants using spurious acoustic information present in the audio files and 2) using different tokens between training and testing ensures that any indexicality effects are driven by voice rather than recording-specific information.

Visual stimuli consisted of color pictures of the eight familiar objects. For each object we selected three different pictures (see Table 2); two were taken from the University of Iowa MACLab stimulus database (Apfelbaum, Blumstein, & McMurray, 2011; McMurray, Samelson, Lee, & Tomblin, 2010) and one was taken from the Multilingual Picture (MultiPic) databank (Duñabeitia et al., 2017). All images were 240  $\times$  240 pixels during presentation.

---

<sup>1</sup> Pitch measurements were extracted from the final set of experimental stimuli (10 recordings per speaker for each of the eight words). This also applies to the child speaker recorded for Experiment 2.

### *Procedure*

Participants were seated in front of a computer screen and were instructed that their task would be to learn a set of new words and their meanings. They were familiarized with each of the eight items by seeing a picture of its referent accompanied by its orthographic label. Participants were explicitly asked to not read the words aloud. After familiarization, they were fitted with an SR Research EyeLink 2K eye-tracker, a system with remote desktop mounting.

After calibration, participants were given instructions for the training phase, and training began. At the beginning of each trial, pictures of four different objects were presented in the four corners of a 19" monitor operating at a resolution of  $1204 \times 768$  pixels. One picture was the target item for that trial. The other three pictures were fillers, each one corresponding to one of the other seven objects (i.e., two pictures of the same object were never presented on the same trial). In each display, two of the pictures always corresponded to items assigned to the indexical condition and the other two corresponded to items assigned to the uncorrelated condition. All pictures were presented the same number of times (84, out of which 21 as the target) and their position was randomized across trials.

Simultaneously with the presentation of the four pictures, a red circle appeared at the center of the screen. After 500 ms, the circle turned green, cueing the participant to click on it to start the trial. This allowed the participants to briefly look at the pictures before hearing anything, thus minimizing eye movements due to visual search (rather than lexical processing). As soon as participants clicked on the green circle, it disappeared and an auditory stimulus was played through high quality headphones. Participants then clicked on the picture they believed to be the referent of the word. After a 330 ms delay, they received auditory feedback (the Windows sounds "Speech On.wav" and "chord.wav" were used for positive and negative feedback respectively). Training took about 35 mins.

Immediately after training, participants started the testing phase. The procedure and task were the same as in training, but during testing participants did not receive any feedback at the end of each trial. Testing took about 15 mins.

Participants had the chance to take a break every 12 trials (both for the training and testing phases). There was no time limit on the trials, however, participants typically responded in less than 2 sec ( $M = 1,703$  ms,  $SD = 248$  ms).

At the end of the experiment, participants were given a short questionnaire assessing their awareness of the experimental manipulation (i.e., regarding voice-picture relationships).

### *Eye-tracking Recording and Analysis*

Participants were calibrated using the standard 9-point display and monocular eye movements were recorded at a sampling rate of 1,000 Hz (but were resampled at 250 Hz during pre-processing). As in previous studies (Kapnoula, Packard, Gupta, & McMurray, 2015; McMurray, Tanenhaus, & Aslin, 2002), this was automatically parsed into saccades and fixations using default psychophysical parameters. Adjacent saccades and fixations were combined into a single “look” that started at the onset of the saccade and ended at the offset of the fixation.

Eye movements were recorded from the onset of the trial (the green circle) through the participants' response (mouse click). This resulted in a variable trial offset time, depending on the individual response time. We adopted the approach of many prior studies (Allopenna et al., 1998; McMurray et al., 2002) by setting a fixed trial duration of 2,000 ms. If a trial ended before this point, we extended the last eye movement; trials longer than 2,000 ms were truncated. This approach assumes that any fixations made in the very late portions of a trial reflect the word the participant settled on and should, thus, be interpreted as an estimate of the final state of the system.

In converting the coordinates of each look to the object being fixated, the boundaries of the regions of interest containing the objects were extended by 100 pixels in order to account for noise and/or head-drift in the eye-track record. This did not result in any overlap between the objects (the dead space between pictures was 8 pixels vertically and 264 pixels horizontally).

### **Results**

Due to technical problems<sup>2</sup>, six participants were excluded from the analyses of fixations (but were included in the analyses of responses).

---

<sup>2</sup> Technical problems either related to the script/output files, or to participants' pattern of looking resulting in insufficient eye-movement data.



### *Analyses of responses*

The behavioral results (reaction times [RT] and accuracy) were assessed to determine whether the participants learned the words successfully. No effect of congruency was found on either of these measures.

*Training.* Participants performed the task without difficulties and responded in a prompt manner. Average accuracy across training was 95.5% (SD = .6%), which corresponds to 23 (out of 504) error trials per participant. Trials with incorrect responses and/or reaction times (RT) over 5,000 ms (average 9 trials per participant) were excluded from RT analyses. In the remaining trials, average RT was 1,588 ms (SD = 170 ms).

We examined the effect of experimental condition on participants' responses during training using paired t-tests. Average accuracy was 95.4% for trials in the indexical condition and 95.6% for those in the uncorrelated condition. No significant difference was found between them<sup>3</sup>,  $t_{(47)} = 0.869$ ,  $p = 0.389$ . Similarly, no significant difference was found in RTs between the indexical (M = 1,587 ms) and uncorrelated (M=1,590 ms) conditions,  $t_{(47)} = -0.142$ ,  $p = 0.887$ .

*Testing.* Average accuracy in testing was 99.5% (SD = .1%), which corresponds to 1 error trial (out of 216) per participant. Any trials with incorrect responses and/or RT over 5,000 ms (average 4 trials per participant) were excluded from RT analyses. In the remaining trials, average RT was 1,481 ms (SD = 253 ms).

We compared performance on items that had been assigned to the indexical versus the uncorrelated conditions during training. Average accuracy was 99.6% for trials from the indexical and 99.4% for trials from the uncorrelated condition,  $t_{(47)} = 1.052$ ,  $p = 0.298$ . Average RT was 1,454 ms for trials from the indexical and 1,456 ms for trials from the uncorrelated condition,  $t_{(47)} = -0.088$ ,  $p = 0.930$ . Next, we tested whether trials from within the indexical condition differed as a function of whether they were congruent or incongruent with the voice-picture associations formed during training. Average accuracy was 99.6% for trials in both conditions,  $t_{(47)} = -0.102$ ,  $p = 0.919$ . Similarly, no significant difference was found in RTs between congruent (M = 1,453 ms) and incongruent (M=1,455 ms) trials,  $t_{(47)} = -0.094$ ,  $p = 0.926$ .

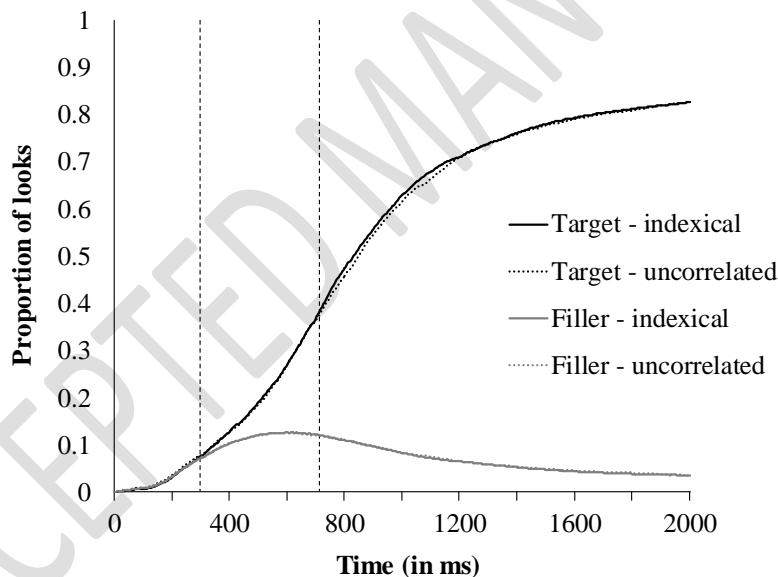
---

<sup>3</sup> All raw accuracy percentages in all experiments were transformed using the empirical logit transformation prior to any analysis.

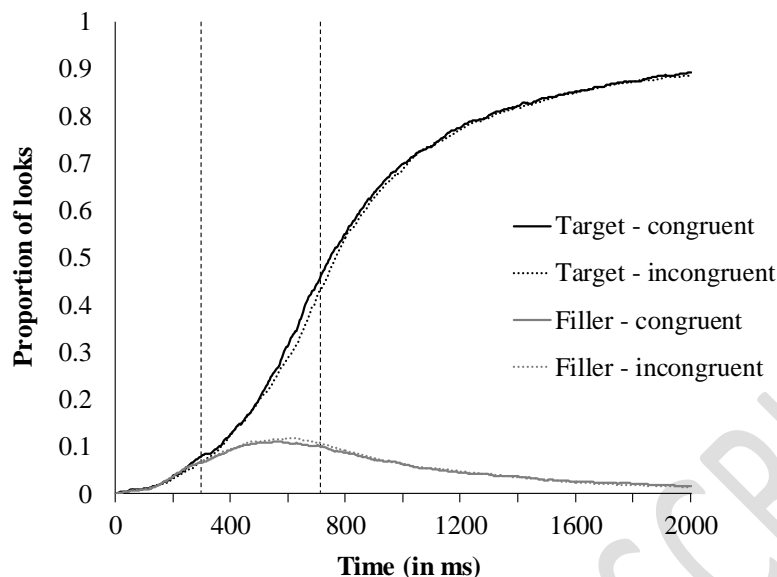
*Analyses of fixations*

Our primary analyses are based on the eye tracking data. We start by reporting the training results, to verify that participants learned the words successfully. Then, we move to the testing results, addressing two questions. First, does congruency significantly affect the successful activation of the target? Second, does congruency allow for faster access to the target's meaning? This second question concerns the core issue of our study: Does indexical information directly affect the time-course of lexical activation?

*Proportion of looks analyses.* For both the training and testing phases, we computed the average proportion of looks to the target picture in the time window corresponding to the duration of the auditory stimulus ( $M = 562$  ms), corrected for the 50 ms of silence at the onset and offset of the auditory stimulus file plus 200 ms oculomotor delay (see Figures 3 and 4).



*Figure 3.* Proportion of looks to the target versus proportion of looks to the fillers (averaged across the three fillers) as a function of condition (indexical versus uncorrelated) during training in Experiment 1. Vertical dotted lines mark the onset and mean offset of the auditory stimuli corrected for the 50 ms of silence at the onset and offset of the auditory stimulus file plus the 200 ms oculomotor delay.



*Figure 4.* Proportion of looks to the target versus proportion of looks to the fillers (averaged across the three fillers) as a function of voice-picture congruency during testing in Experiment 1. Vertical dotted lines mark the onset and mean offset of the auditory stimuli corrected for the 50 ms of silence at the onset and offset of the auditory stimulus file plus the 200 ms oculomotor delay.

No significant difference was found between the indexical and uncorrelated conditions during training<sup>4</sup>,  $t_{(41)} = 0.342$ ,  $p = 0.734$  (see Figures 3 and 5.A). We then evaluated the effect of congruency during testing and found a non-significant trend<sup>5</sup>,  $t_{(41)} = 1.993$ ,  $p = 0.053$ : There was a slightly higher proportion of looks to the target picture for the congruent ( $M = 27.2\%$ ) compared to the incongruent trials ( $M = 25.5\%$ ; see Figures 4 and 5.B).

<sup>4</sup> All raw fixation proportions in all experiments were transformed using the logit transformation prior to any analysis.

<sup>5</sup> Which was significant according to the one-tail test,  $p = 0.026$ . Given that only a difference favoring the congruent condition would make sense, the one-tailed test is arguably appropriate.

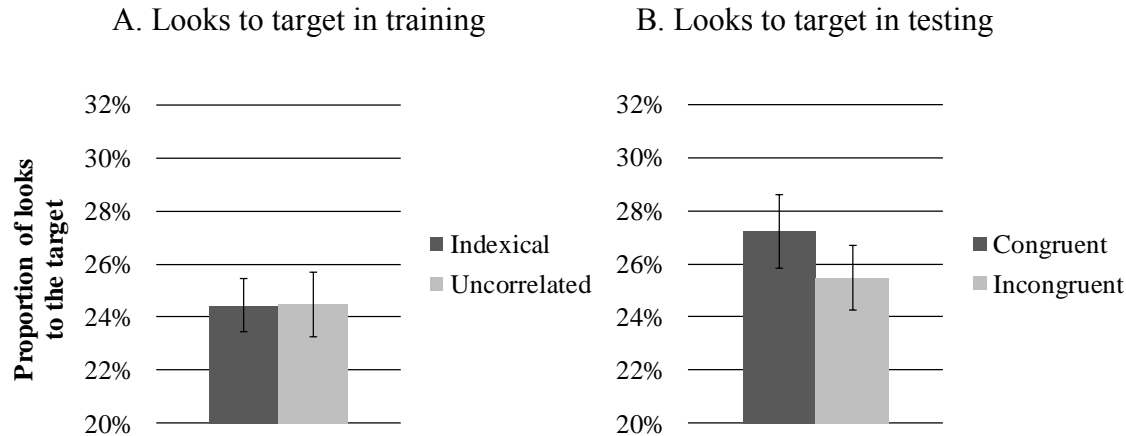


Figure 5. Average proportions of looks to the target in training and testing in Experiment 1. The error bars represent standard errors. Note: See Figure S.1 in Supplemental Materials for the corresponding differences in proportions of looks to target across the three Experiments.

*Curve-fitting analyses.* The marginally greater proportion of looks to the target in the congruent condition (significant by the one-tailed test; see Footnote 5) is in line with the idea that referent selection may have been facilitated by congruency. However, the proportion statistic is a rather coarse measure. To assess whether there were significant differences in the *dynamic build-up* of target activation, we evaluated differences between time-courses using a nonlinear curve-fitting approach (Farris-Trimble & McMurray, 2013; McMurray et al., 2010; Seedorff, Oleson, & McMurray, 2018). This approach avoids the limitations of analyses that are restricted to a specific time window. We fit a four-parameter logistic function (see Eq.1. in McMurray et al., 2010) to the data as a function of time (see Figure 6). In this equation, the lower and higher asymptotes correspond to the baseline and peak of the curve (i.e., minimum and maximum lexical activation, respectively), the slope reflects the speed with which activation grows in time, and the crossover corresponds to the point in time when activation crosses from the lower half of the range to the higher half of the range (e.g., if baseline is 0 and peak is 1, then the crossover would correspond to the point in time when activation is .5). Conceptually, the baseline and peak parameters are static, whereas the slope and crossover parameters capture the dynamics of activation.

In the interest of being comprehensive, we analyzed and report all four parameters. However, since we were mainly interested in the dynamic build-up of the target word's activation, we

focus on the two latter parameters (slope and crossover), which we expected to be more sensitive to the effect of congruency.

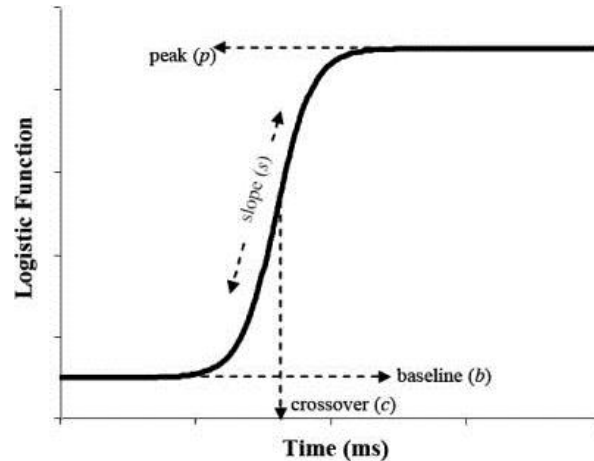


Figure 6. Logistic function parameters: Baseline ( $b$ ; i.e., lower asymptote), peak ( $p$ ; i.e., upper asymptote), slope of the transition ( $s$ ), and crossover point ( $c$ ). The figure was taken from McMurray et al., 2010).

All curve-fitting analyses were implemented using the *bdots* R package (Oleson, Cavanaugh, McMurray, & Brown, 2015; Seedorff, Oleson, Brown, Cavanaugh, & McMurray, 2017). Note that *bdots* can be used to fit non-linear curves (e.g., individuals' looks to a target item in time) and to evaluate differences among time series curves with p-value adjustment for multiple comparisons. Given that in this study we were mainly interested in the dynamic aspects of lexical activation (i.e., how lexical activation unfolds over time, rather than overall higher/lower activation at specific time windows), we opted for using only the former *bdots* tool.

For the curve-fitting analyses, we first computed the average proportions of looks to the target along the entire time-course of the trial (i.e., 0 – 2,000 ms) for each subject and each congruency condition (see Figure 4). Then, we used the *bdots* *logistic.fit* function to find the four-parameter logistic function that provided the best fit for each curve. Overall fits were good<sup>6</sup>. Using the

<sup>6</sup> Average  $R^2$  was 0.99. With the exception of one subject, who was dropped due to failure to fit one of their curves, all remaining (82) curves had good fits ( $R^2 > 0.96$ ). Two curves were fitted without using the default assumption of auto-correlated errors (AR1). According to this assumption, the variance at a time point is correlated with the variance at the time point immediately prior, with the correlation decaying exponentially as time points get further

bdots logistic.boot function (for paired data), we tested the effect of congruency on each of the four parameters (baseline, peak, slope, and crossover; see Seedorff, Oleson, Cavanaugh, & McMurray, 2017; Seedorff et al., 2018, for a presentation of the bdots package, its conceptual implementation, and a discussion of the statistical approach).

Table 3. Comparisons for the four logistic parameters used to describe target activation in the testing phase of Experiment 1

| Parameter                  | Difference between conditions<br>(congruent minus incongruent) | $t_{(40)}$ | SE     | p     |
|----------------------------|--|------------|--------|-------|
| Baseline ( <i>b</i> )      | -0.029   | -1.182     | 0.024  | 0.244 |
| Peak location ( <i>p</i> ) | 0.006  | 1.237      | 0.005  | 0.223 |
| Slope ( <i>s</i> )         | 0.00005  | 3.601      | <0.001 | <.001 |
| Crossover ( <i>c</i> )     | -39.388  | -2.443     | 16.543 | 0.019 |

*Note.* Parameter estimates rely on simulations. All *t* tests are calculated using the bootstrapped means and are adjusted to account for the additional variance around the parameter estimate (Seedorff, Oleson, Cavanaugh, et al., 2017; Seedorff et al., 2018). This means that we are able to account for the uncertainty presented in estimating the parameters (i.e., the subject-specific standard errors around each parameter mean), whereas this information is lost when a traditional *t*-test is performed on the extracted parameter estimates. This note also applies to Tables 4 and 5. *Note 2.* See S2 in Supplemental Materials for a further explanation of the slope values.

The results, shown in Table 3, clearly point to a dynamic advantage for congruency: The curve describing looks to the target grew at a significantly faster rate for congruent than incongruent trials. In addition, the crossover point was significantly earlier for congruent compared to incongruent trials (~39ms). Simply put, trials that maintained the association between a particular voice and a particular picture led to faster access to the referent. The static parameters (baseline and peak) were not affected by the congruency manipulation.

---

apart. This is taken into account during curve-fitting, when computing the error in the process of minimizing the sum of the squared errors between the fitted and observed values.

### ***Questionnaire***

Participants' responses in the questionnaire administered after the experiment confirmed our assumption that participants were not aware of our manipulation of voice-picture relationships. Specifically, no one thought that the talker/voice was part of the manipulation except one participant who mentioned the tone of the voice as a possible key aspect of the experiment (most likely referring to intonation differences between the audio tokens). We will discuss the implications of this lack of explicit knowledge in the General Discussion.

### **Discussion**

The results of Experiment 1 show that indexical information – in this case, a talker's voice – can directly affect referent selection.

It is important to note that in contrast to previous studies, the participants' task was irrelevant to the particular voice in which a target word was spoken, or the particular picture of the referent. This dissociation prevents participants from using indexical information in an explicit manner. Participants' accuracy and response times on the task they were given was not affected by our congruency manipulation. Similarly, overall activation of the target word (as reflected by the probability of fixating the referent) was only marginally lower in incongruent compared to congruent trials. The critical effect of the congruity manipulation was on the time-course of target activation: Participants were faster to access the referent in the congruent compared to the incongruent condition. This pattern indicates a direct link between indexical information and lexical meaning. This is the first study to show that *indexical information can directly affect access to lexical meaning* in the absence of explicitly generated inferences.

Given the importance of replication, Experiment 2 is almost identical to Experiment 1. However, instead of using three adult speakers, we used two adults and one child speaker. This change was meant to test whether the indexical effect would become bigger if the three speakers were more distinguishable from each other, while also testing for the replicability of the core results.

## **Experiment 2: Are Lexical Indexical Effects Graded?**

### **Method**

#### ***Participants***

Forty-eight (33 females; mean age = 23.9 years) native speakers of Spanish participated in Experiment 2. As in Experiment 1, most participants were also fluent in Basque. All participants had normal/corrected-to-normal vision and no known hearing or neurological impairments. Participants underwent informed consent and were remunerated for their participation. All experimental procedures were approved by the BCBL ethics committee.

#### ***Design***

The experimental design was identical to that of Experiment 1.

#### ***Materials***

The set of novel words was the same as in Experiment 1 (see Table 1).

For the auditory stimuli, we used the stimuli recorded by the two adult speakers from Experiment 1 with the lowest pitches (one male and one female). Stimuli spoken by the second female speaker (with the highest pitch) were replaced by new recordings spoken by a young (11 y.o.) girl (pitch:  $M = 234.5\text{Hz}$ ,  $SD = 6.4$ ). All new stimuli were recorded in the same room and at the same sampling rate as the stimuli used in Experiment 1. All stimulus editing and preparation procedures were identical to those of Experiment 1.

Visual stimuli were the same as in Experiment 1.

#### ***Procedure***

All procedures were identical to those of Experiment 1.

#### ***Eye-tracking Recording and Analysis***

All eye-tracking recording and analysis protocols were identical to those of Experiment 1.

### **Results**

Due to technical problems, five participants were excluded from the analyses of fixations (but were included in the analyses of responses).



### *Analyses of responses*

As in Experiment 1, we analyzed the behavioral results to verify that participants had successfully learned the words. No effect of congruency was found on either accuracy or reaction time.

*Training.* Participants performed the task without difficulties and responded in a prompt manner. Average accuracy across training was 94.4% (SD = 7%), which corresponds to 12 (out of 504) error trials per participant. Trials with incorrect responses and/or RT over 5,000 ms (average 10 trials per participant) were excluded from RT analyses. For the remaining trials, average RT was 1,579 ms (SD = 237 ms).

We examined the effect of experimental condition on participants' responses during training using paired t-tests. Average accuracy was 94.7% for trials in the indexical condition and 94.6% for those in the uncorrelated condition. No significant difference was found between them,  $t_{(47)} = 0.198$ ,  $p = 0.844$ . However, participants were significantly faster in the indexical ( $M = 1,559$  ms) compared to the uncorrelated ( $M = 1,600$  ms) condition,  $t_{(47)} = -2.111$ ,  $p = 0.040$ .

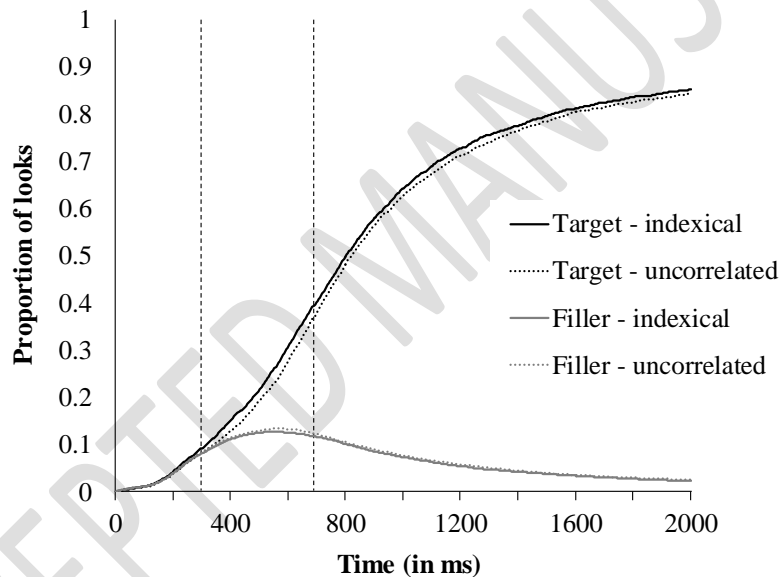
*Testing.* Average accuracy in testing was 98.7% (SD = 3%), which corresponds to 3 error trials (out of 216) per participant. Any trials with incorrect responses and/or RT over 5,000 ms (average 4 trials per participant) were excluded from RT analyses. For the remaining trials, average RT was 1,426 ms (SD = 230 ms).

We compared performance on items that had been assigned to the indexical versus the uncorrelated conditions during training. Average accuracy was 98.8% for trials from the indexical condition and 98.6% for trials from the uncorrelated condition,  $t_{(47)} = 1.008$ ,  $p = 0.319$ . Average RT was 1,407 ms for trials from the indexical and 1,444 ms for trials from the uncorrelated condition,  $t_{(47)} = -2.093$ ,  $p = 0.042$ . Next, we tested whether trials from within the indexical condition differed as a function of whether they were congruent or incongruent with the voice-picture associations formed during training. Average accuracy was 98.8% for trials in both conditions,  $t_{(47)} = .112$ ,  $p = 0.911$ . Similarly, no significant difference was found in RTs between congruent ( $M = 1,403$  ms) and incongruent ( $M = 1,409$  ms) trials,  $t_{(47)} = -0.378$ ,  $p = 0.707$ .

*Analyses of fixations*

As in Experiment 1, we examined participants' looks in two ways: First we asked whether there were any differences in the probability of fixating the target, and then we looked at the dynamic build-up of target activation following a curve-fitting approach, implemented using the *bdots R* package (Oleson et al., 2015).

*Proportion of looks analyses.* We computed the average proportion of looks to the target picture in the time window corresponding to the duration of the auditory stimulus ( $M = 543$  ms), corrected for the 50 ms of silence at the onset and offset of the auditory stimulus file plus the 200 ms oculomotor delay (see Figures 7 and 8).



*Figure 7.* Proportion of looks to the target versus proportion of looks to the fillers (averaged across the three fillers) as a function of condition (indexical versus uncorrelated) during training in Experiment 2. Vertical dotted lines mark the onset and mean offset of the auditory stimuli corrected for the 50 ms of silence at the onset and offset of the auditory stimulus file plus the 200 ms oculomotor delay.

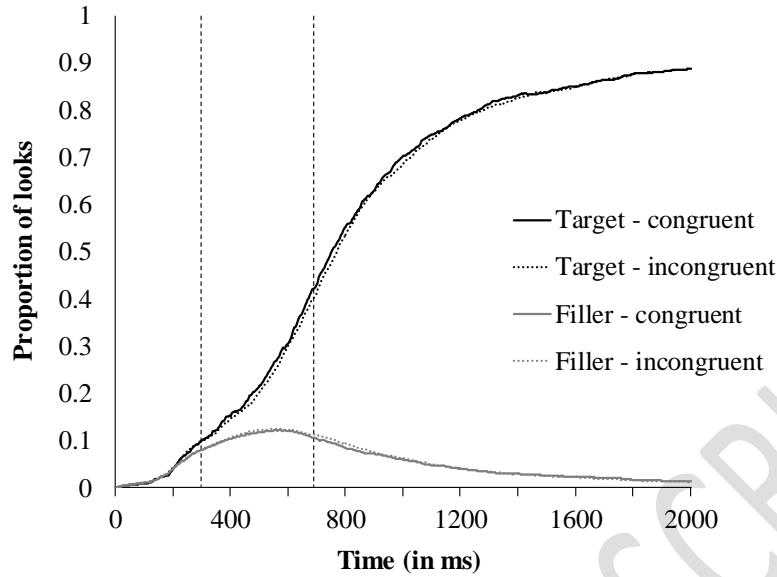
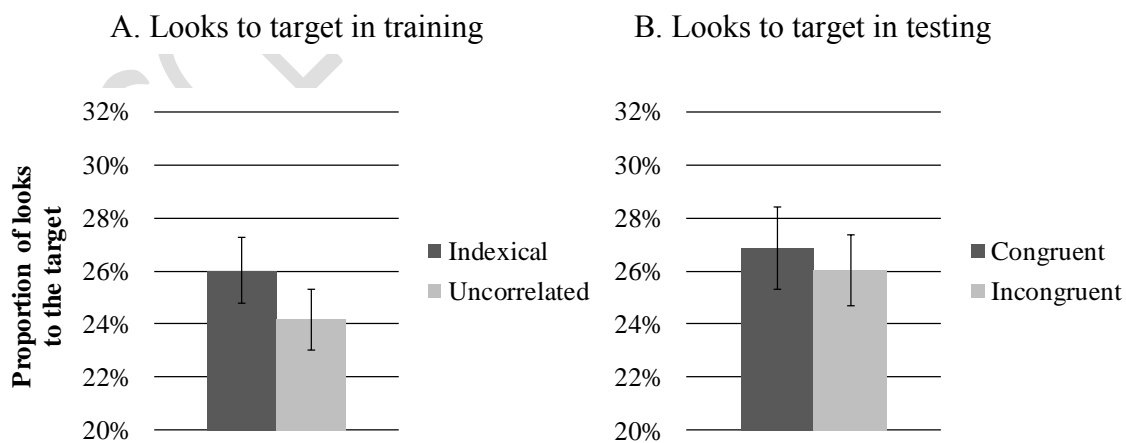


Figure 8. Proportion of looks to the target versus proportion of looks to the fillers (averaged across the three fillers) as a function of voice-picture congruency during testing in Experiment 2. Vertical dotted lines mark the onset and mean offset of the auditory stimuli corrected for the 50 ms of silence at the onset and offset of the auditory stimulus file plus the 200 ms oculomotor delay.

A non-significant numerical difference<sup>7</sup> was found between indexical and uncorrelated conditions in training,  $t_{(42)} = 1.948$ ,  $p = 0.058$  (see Figures 7 and 9.A). This difference remained non-significant in testing,  $t_{(42)} = 1.122$ ,  $p = 0.268$ . Finally, no congruency effect was found in testing,  $t_{(42)} = -0.077$ ,  $p = 0.939$  (see Figures 8 and 9.B).



<sup>7</sup> Which was significant according to the one-tail test,  $p = 0.029$ .

*Figure 9.* Average proportions of looks to the target in training and testing in Experiment 2. The error bars represent standard errors. Note: See Figure S.1 in Supplemental Materials for the corresponding differences in proportions of looks to target across the three Experiments.

As in Experiment 1, we next proceeded to a more fine-grained analysis of participants' eye movements, focusing on the dynamic build-up of target activation.

*Curve-fitting analyses.* For the curve-fitting analyses, we first computed average proportions of looks to the target along the entire time-course of the trial (i.e. 0 – 2,000 ms) for each subject and each congruency condition (see Figures 7 and 8). Then, we used the `bdots logistic.fit` function to find the four-parameter logistic function that provided the best fit for each curve. Overall fits were good<sup>8</sup>.

We next used the `bdots logistic.boot` function (for paired data) to examine the effect of congruency on each of the four parameters (baseline, peak, slope, and crossover).

Table 4. Comparisons for the four logistic parameters used to describe target activation in the testing phase of Experiment 2

| Parameter                  | Difference between conditions<br>(congruent minus incongruent) | $t_{(42)}$ | SE     | p     |
|----------------------------|--|------------|--------|-------|
| Baseline ( <i>b</i> )      | -0.024   | -1.114     | 0.022  | 0.272 |
| Peak location ( <i>p</i> ) | 0.002  | 0.788      | 0.003  | 0.435 |
| Slope ( <i>s</i> )         | 0.000002   | 0.106      | <0.001 | 0.916 |
| Crossover ( <i>c</i> )     | -28.773  | -2.095     | 14.017 | 0.042 |

As seen in Table 4, the crossover point of the curve describing looks to the target was significantly earlier for congruent compared to incongruent trials. In line with the results of Experiment 1, this result suggests that the target word was activated faster on congruent than on

<sup>8</sup> Average  $R^2$  was 0.99. All 86 curves had good fits ( $R^2 > 0.95$ ). Three curves were fitted without AR1.

incongruent trials. The crossover difference (~29ms) reflects an estimate of that difference. Unlike Experiment 1, the dynamic difference was limited to the crossover parameter, with no effect seen for the slope parameter. As in Experiment 1, there were no reliable effects on the two static parameters.

## **Discussion**

The results of Experiment 2 match those of Experiment 1 in showing that indexical information can directly affect referent selection. As in Experiment 1, participants' accuracy and response times were not affected by congruency. In contrast to Experiment 1, overall activation of the target word (as reflected by the probability of fixating the referent) did not differ between congruent and incongruent trials, indicating a somewhat weaker and less stable effect. In Experiment 1, we determined that congruency led to faster access to meaning, with significant differences in both the slope and crossover point. In the current experiment, we again found a difference in dynamic activation, but this difference was limited to the crossover parameter.

Contrary to our expectation, Experiment 2 did not lead to a stronger effect compared to Experiment 1, despite our using three talker voices that were acoustically more distinct. Using the crossover difference as the metric, the target words were activated 29 ms faster in the congruent compared to the incongruent condition in Experiment 2, whereas this difference was 39 ms in Experiment 1. The similar-sized effect could mean that voice information is encoded in terms of talker identity rather than in a continuous way, i.e., three distinguishable voices are treated the same way, regardless of how distinguishable they are. Alternatively, it might mean that the pitch difference of 40 Hz between the swapped talkers used in the two experiments was not big enough to lead to a noticeable difference. While this question remains open, the results of Experiments 1 and 2 collectively clearly demonstrate that indexical information can directly affect access to lexical meaning.

## **Experiment 3: Do Lexical Indexical Effects Change After 24 Hours?**

Even though sleep may not be a prerequisite for the integration of novel words into the mental lexicon (Coutanche & Thompson-Schill, 2014; Fernandes, Kolinsky, & Ventura, 2009; Kapnoula & McMurray, 2016; Kapnoula et al., 2015), there is strong evidence that certain aspects of

lexical representations/processes are enhanced through sleep-based consolidation. Notably, sleep-based consolidation has been found to promote integration of novel word-forms (Davis & Gaskell, 2009; Dumay & Gaskell, 2007, 2012; Tamminen, Payne, Stickgold, Wamsley, & Gaskell, 2010). In Experiment 3, we investigate whether sleep might modulate the indexical effect observed in Experiments 1 and 2.

On the one hand, if we conceptualize words as multi-dimensional vectors (Gaskell & Marslen-Wilson, 1997; Goldinger, 1998), then voice information can be conceived of as one of the dimensions of that vector. According to this view, sleep-based consolidation should stabilize and strengthen all aspects of novel words, including this dimension. As a result, sleep may boost the effect observed in Experiments 1 and 2.

On the other hand, according to the Complementary Learning Systems Account (McClelland, McNaughton, & O'Reilly, 1995), offline consolidation leads to the integration of experience with prior knowledge, via the abstraction of certain features away from episodic information (Xie, Earle, & Myers, 2018). In other words, sleep is thought to promote the formation of abstracted representations. In contrast to the experimental manipulation employed here, in typical situations, voice information is not a relevant dimension for lexical access. As a result, the system may have been tuned to abstract away such information. According to this rationale, sleep should weaken the effect observed in Experiments 1 and 2.

In Experiment 3 we introduced a 24-hr delay between training and testing to examine the effect of sleep on the observed congruency effect.

## **Method**

### ***Participants***

Fifty (40 females; mean age = 23.8 years) native speakers of Spanish participated in Experiment 3. As in Experiments 1 and 2, most participants were also fluent in Basque. All participants had normal/corrected-to-normal vision and no known hearing or neurological impairments.

Participants underwent informed consent and were remunerated for their participation. All experimental procedures were approved by the BCBL ethics committee.

### ***Design***

The experimental design was similar to that of Experiments 1 and 2. The critical difference was that the *training* and *testing* phases took place exactly 24 hrs apart.

### ***Materials***

For both the training and the testing phases, all auditory and visual stimuli were the same as in Experiment 1.

### ***Procedure***

All procedures were identical to those of Experiments 1 and 2, with the exception of the 24-hr delay between training and testing.

### ***Eye-tracking Recording and Analysis***

All eye-tracking recording and analysis protocols were identical to those of Experiments 1 and 2.

### **Results**

Due to problematic eye-tracks, four participants were excluded from the analyses of fixations (but were included in the analyses of responses). In addition, the eye-movement results from two participants were excluded from some analyses<sup>9</sup> due to technical problems.

### ***Analyses of responses***

Once again, we analyzed the behavioral results to verify that participants had successfully learned the words. No effect of congruency was found on either accuracy or reaction times.

*Training.* Participants performed the task without difficulties and responded in a prompt manner. Average accuracy across training was 95.0% (SD = 5%), which corresponds to 11 (out of 504) error trials per participant. Trials with incorrect responses and/or RT over 5,000 ms (average 4 trials per participant) were excluded from RT analyses. In the remaining trials, average RT was 1,573 ms (SD = 189 ms).

We examined the effect of experimental condition on participants' responses during training using paired t-tests. Average accuracy was 95.3% for trials in the indexical condition and 95.0%

---

<sup>9</sup> Two eye-movement data files were missing/corrupted (one participant's training data and another participant's testing data). The second participant also had problematic eye-track, leading us to exclude their training data as well.

for those in the uncorrelated condition. No significant difference was found between them,  $t_{(48)} = 0.604$ ,  $p = 0.549$ . Similarly, no significant difference was found in RTs between indexical ( $M = 1,567$  ms) and uncorrelated ( $M=1,581$  ms) conditions,  $t_{(48)} = -0.635$ ,  $p = 0.529$ .

*Testing.* Average accuracy in testing was 99.5% ( $SD = 1\%$ ), which corresponds to 1 error trial (out of 216) per participant. Any trials with incorrect responses and/or RT over 5,000 ms (average 1 trial per participant) were excluded from RT analyses. For the remaining trials, average RT was 1,345 ms ( $SD = 154$  ms).

We compared performance on items that had been assigned to the indexical versus the uncorrelated conditions during training. Average accuracy was 99.5% for trials from the indexical and 99.4% for trials from the uncorrelated condition,  $t_{(48)} = 0.332$ ,  $p = 0.741$ . Average RT was 1,332 ms for trials from the indexical and 1,355 ms for trials from the uncorrelated condition,  $t_{(48)} = -1.252$ ,  $p = 0.217$ . Next, we tested whether trials from within the indexical condition differed as a function of whether they were congruent or incongruent with the voice-picture associations formed during training. Average accuracy was 99.5% for trials in both conditions,  $t_{(48)} = -0.088$ ,  $p = 0.930$ . Similarly, no significant difference was found in RTs between congruent ( $M = 1,326$  ms) and incongruent ( $M=1,335$  ms) trials,  $t_{(48)} = -0.690$ ,  $p = 0.493$ .

### *Analyses of fixations*

As in Experiments 1 and 2, we first asked whether there were any differences in the probability of fixating the target and then looked at the dynamic build-up of target activation following a curve-fitting approach, implemented using the *bdots* R package (Oleson et al., 2015).

*Proportion of looks analyses.* We computed the average proportion of looks to the target picture in the time window corresponding to the duration of the auditory stimulus ( $M = 562$  ms), corrected for the 50 ms of silence at the onset and offset of the auditory stimulus file plus the 200 ms oculomotor delay (see Figures 10 and 11).



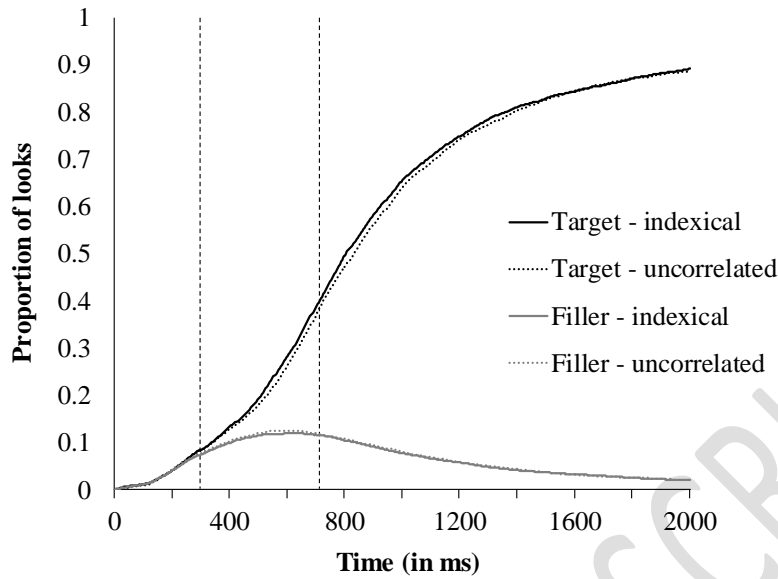


Figure 10. Proportion of looks to the target versus proportion of looks to the fillers (averaged across the three fillers) as a function of condition (indexical versus uncorrelated) during training in Experiment 3. Vertical dotted lines mark the onset and mean offset of the auditory stimuli corrected for 200 ms oculomotor delay plus 50 ms for silence added to stimulus onset.

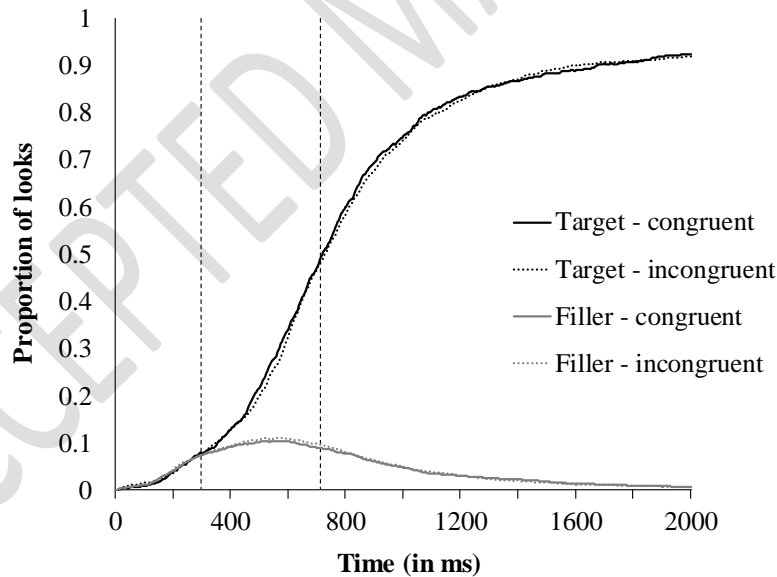


Figure 11. Proportion of looks to the target versus proportion of looks to the fillers (averaged across the three fillers) as a function of voice-picture congruency during testing in Experiment 3. Vertical dotted lines mark the onset and mean offset of the auditory stimuli corrected for 200 ms oculomotor delay plus 50 ms for silence added to stimulus onset.

There was no significant difference between indexical and uncorrelated conditions in training,  $t_{(43)} = 1.548$ ,  $p = 0.129$  (see Figures 10 and 12.A), or testing,  $t_{(44)} = 0.207$ ,  $p = 0.837$ , and no congruency effect was found in testing,  $t_{(44)} = 0.496$ ,  $p = 0.622$  (see Figures 11 and 12.B).

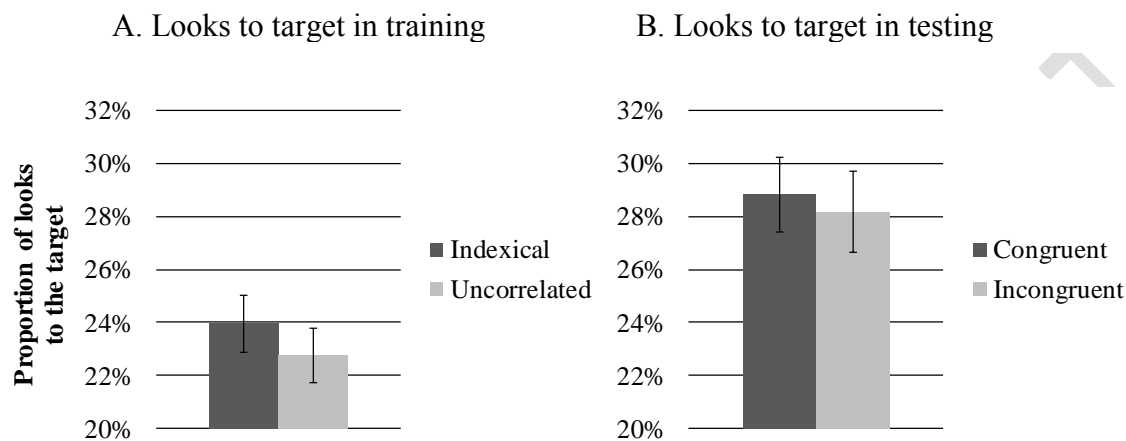


Figure 12. Average proportions of looks to the target in training and testing in Experiment 3. The error bars represent standard errors. Note: See Figure S.1 in Supplemental Materials for the corresponding differences in proportions of looks to target across the three Experiments.

We next examined the time-course of target activation using curve-fitting.

*Curve-fitting analyses.* For the curve-fitting analyses, we first computed average proportions of looks to the target along the entire time-course of the trial (i.e. 0 – 2,000 ms) for each subject and each congruency condition (see Figures 10 and 11). We used the `bdots logistic.fit` function to find the four-parameter logistic function that provided the best fit for each curve. Overall fits were good<sup>10</sup>.

Next, we used the `bdots logistic.boot` function (for paired data) to examine the effect of congruency on each of the four parameters (baseline, peak, slope, and crossover).

<sup>10</sup> Average  $R^2$  was 0.99. With the exception of one subject, who was dropped due to poor fitting in one of their curves, all remaining (88) curves had good fits ( $R^2 > 0.95$ ). Three curves were fitted without AR1.

Table 5. Comparisons for the four logistic parameters used to describe target activation in the testing phase of Experiment 3

| Parameter             | Difference between conditions | $t_{(43)}$ | SE     | p     |
|-----------------------|-------------------------------|------------|--------|-------|
| Baseline ( $b$ )      | -0.008                        | -0.950     | 0.008  | 0.347 |
| Peak location ( $p$ ) | 0.003                         | 1.085      | 0.002  | 0.284 |
| Slope ( $s$ )         | 0.00003                       | 2.094      | <0.001 | 0.042 |
| Crossover ( $c$ )     | -11.759                       | -2.803     | 4.171  | 0.008 |

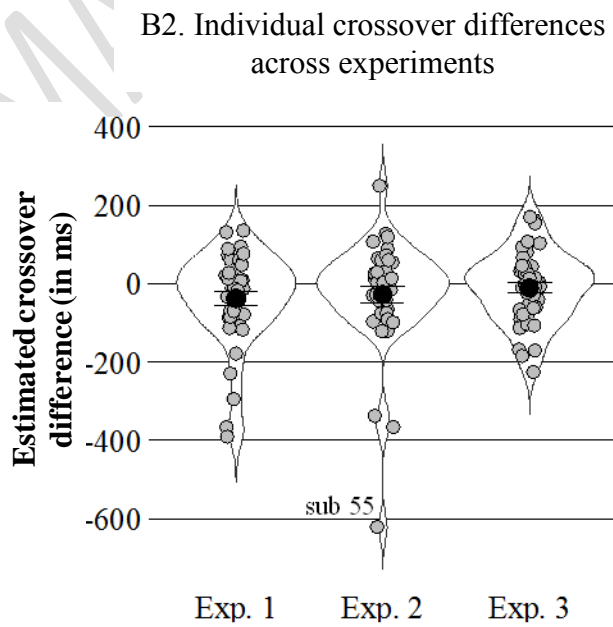
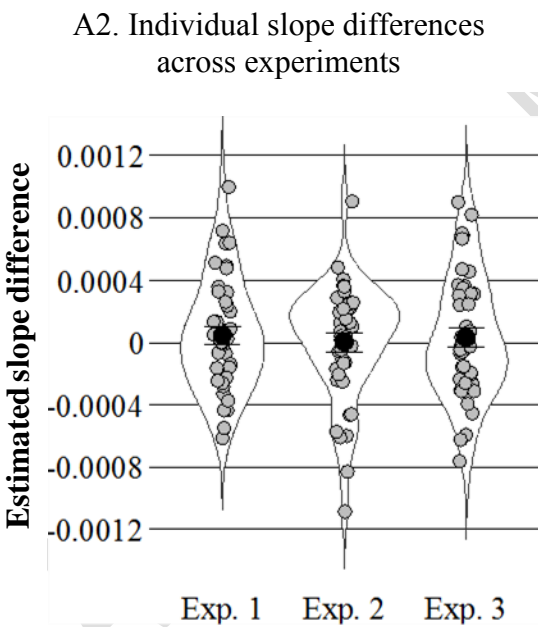
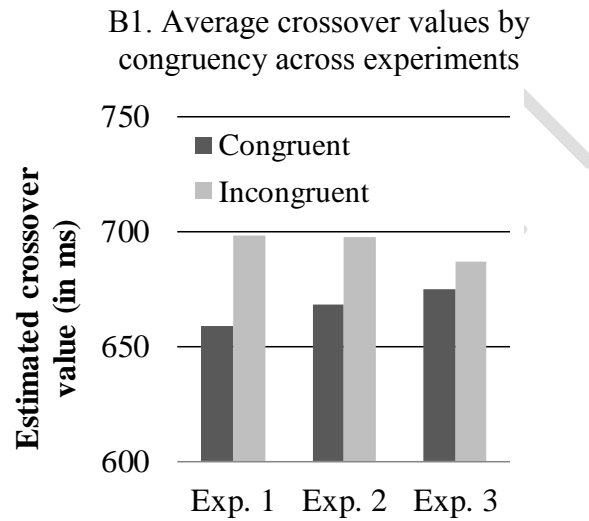
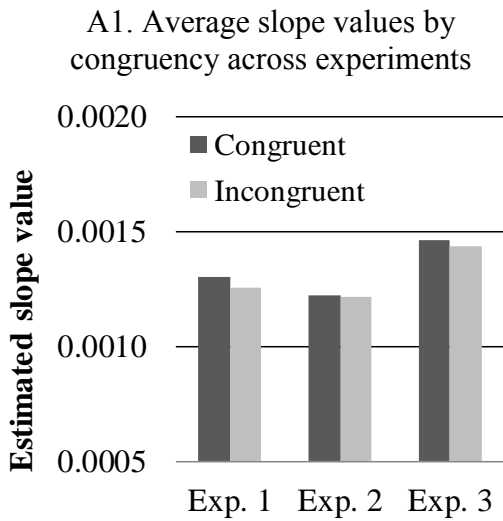
As seen in Table 5, the curve describing looks to the target grew at a significantly faster rate for congruent than incongruent trials. In addition, the crossover point was significantly earlier for congruent compared to incongruent trials. The two static parameters showed no effects. Thus, as in Experiments 1 and 2, the target word was activated faster in congruent than incongruent trials. This dynamic difference produced significant differences in the slope parameter of Experiments 1 and 3, and significant differences in the crossover parameter in all three experiments. The estimate of the dynamic difference in Experiment 3, in terms of the the crossover parameter difference, is ~12 ms.

To examine whether this effect was modulated by sleep, we conducted two  $2 \times 3$  ANOVAs with Congruency as a within-subject factor and Experiment as a between-subjects factor. One ANOVA used slope as the dependent variable, and the second used crossover as the dependent variable<sup>11</sup>. Figure 13 summarizes the effects being tested in these analyses (see Table S.1 in Supplemental Materials for the detailed ANOVA report).

For the slope, there was no main effect of Congruency,  $F_{(1, 125)} = 1.536$ ,  $p = 0.218$ ,  $\eta^2 = .012$ , a marginally significant main effect of Experiment,  $F_{(2, 125)} = 2.774$ ,  $p = .066$ ,  $\eta^2 = .043$ , and no Congruency  $\times$  Experiment interaction,  $F_{(2, 125)} = 0.720$ ,  $p = 0.489$ ,  $\eta^2 = .011$ .

<sup>11</sup> Parameter estimates were extracted from the *bdots* logistic.fit function output (Seedorff, Oleson, Cavanaugh, et al., 2017; Seedorff et al., 2018) and ANOVAs were performed in SPSS.

For the crossover, there was a main effect of Congruency,  $F_{(1, 125)} = 6.822$ ,  $p = .010$ ,  $\eta^2 = .052$ , no main effect of Experiment,  $F_{(2, 125)} = 0.027$ ,  $p = .973$ ,  $\eta^2 = .000$ , and no Congruency  $\times$  Experiment interaction,  $F_{(2, 125)} = 0.689$ ,  $p = .504$ ,  $\eta^2 = .011$ .



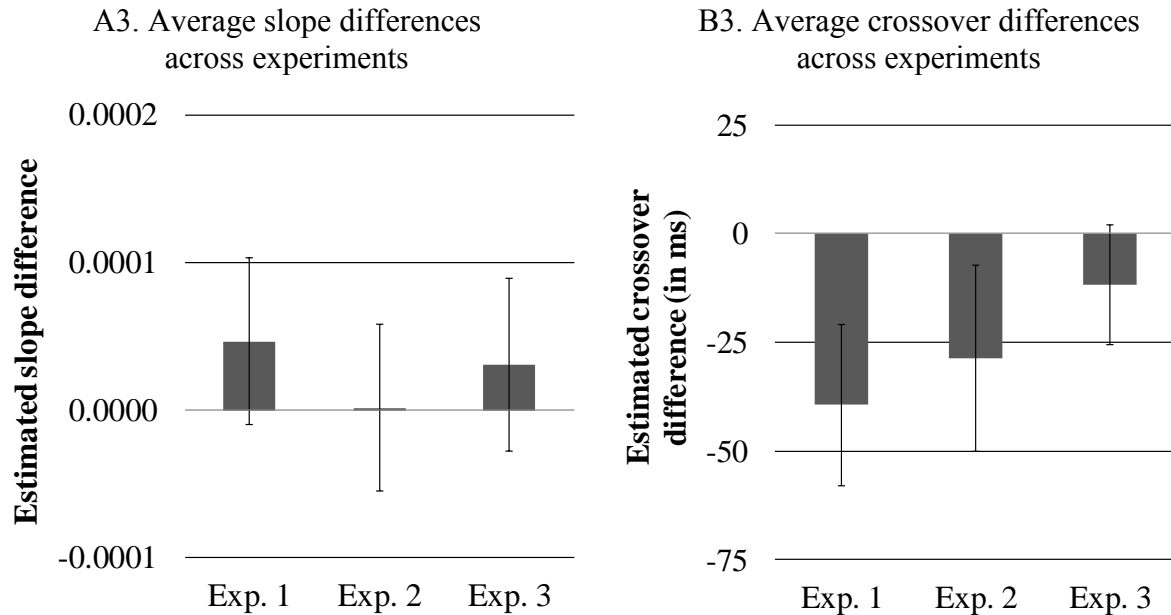


Figure 13. Estimated slope and crossover parameters by Congruency across Experiments. The error bars represent standard errors of the difference scores. Note: When subject 55 (see far low data point in panel B2) is removed, main effect of Congruency remains significant,  $F_{(1, 124)} = 5.790$ ,  $p = 0.018$ ,  $\eta^2 = 0.045$ .

The absence of significant interactions reflects the fact that the effect of congruency was comparable across experiments.

## Discussion

The results of Experiment 3 replicate those of Experiments 1 and 2 in demonstrating that indexical information can have a direct effect on spoken word recognition all the way to referent selection. As in the two previous experiments, we found a significant effect of congruency on the crossover parameter and, as in Experiment 1, we found a significant effect of congruency on the slope parameter as well, reflecting faster activation of the target word when the voice information matched the visual referent.

The effect of congruency on the crossover point was smaller in Experiment 3 (about a ~12 ms crossover difference in Experiment 3 compared to ~39 and ~29 ms in Experiments 1 and 2). The direction of this effect is in line with the second alternative presented earlier, according to which, sleep supports the formation of abstracted representations that are stripped off of typically irrelevant information (such as talker information). However, our analyses showed that this decrease was not statistically significant (i.e., the effect of congruency was not modulated by the

presence of sleep, as indicated by the non-significant Congruency  $\times$  Experiment interactions). This pattern of results is consistent with the view that freshly learned words can be adequately integrated into the mental lexicon both at the word-form (Coutanche & Thompson-Schill, 2014; Fernandes et al., 2009; Kapnoula & McMurray, 2016; Kapnoula et al., 2015; Lindsay & Gaskell, 2013) and at the semantic level (Borovsky, Elman, & Kutas, 2012; Borovsky, Kutas, & Elman, 2010).

Another point that is worth mentioning is the pattern for the slope parameter across Experiments, a parameter that reflects the speed of target activation. As shown in Figure 13, there is a numerically higher slope estimate for Experiment 3, indicating a sharper slope after a period of sleep. This effect was not significant, but its direction is consistent with the growing literature on the supportive role of sleep in regards to the integration of novel word-forms into the mental lexicon (Davis & Gaskell, 2009; Dumay & Gaskell, 2007, 2012; Tamminen et al., 2010).

### Supplementary Analyses

An alternative interpretation of the results from these three experiments is that participants simply learned associations between voices and images. To investigate this possibility, we conducted a set of additional analyses that are reported in detail in the Supplemental Materials.

We first assessed whether participants were more distracted by filler pictures that were associated with the voice heard in a given trial. Testing a number of measures, we found one marginally significant effect in the expected direction (i.e., numerically higher distraction from voice-congruent fillers;  $F_{(1, 110)} = 3.876$ ,  $p = .052$ ,  $\eta^2 = .034$ ). We then assessed whether this difference was correlated with the main congruency effect. It was not, regardless of whether the two effects were computed from different trials,  $r_{(111)} = 0.093$ ,  $p = 0.330$ , or if they derived from same trials,  $r_{(98)} = -0.066$ ,  $p = 0.710$ .

Together, these results indicate that independently of whether listeners may (or may not) be able to form associations between voices and images, such associations do not seem to be the driving force behind the congruency effect observed in our three experiments.

## General Discussion

We investigated whether indexical information about a talker's voice can affect access to a word's meaning (i.e., referent selection) in a direct and automatic way. We found evidence for a positive answer to this question across all three experiments. This is the first demonstration of such a connection between indexical information and access to meaning.

Following previous work (Creel & Tumlin, 2011), we used a novel-word learning paradigm that enabled us to manipulate the indexical information present in the speech signal and the way in which this information is linked to the word's referent. Critically, in contrast to previous studies, we used an experimental design that prevented participants from explicitly using voice information to access lexical meaning. After exposing participants to specific patterns of word-to-voice-to-referent co-occurrence, we evaluated their sensitivity to these co-occurrences. Our main analyses were focused on the build-up of lexical activation when the indexical information matched the previously learned pattern of co-occurrences (congruent trials) versus when it did not (incongruent trials). These analyses show that participants were faster at accessing lexical meaning when the voice information had been linked with the word's referent. As we will expand upon below, this finding is in line with an episodic view of the mental lexicon, according to which indexical information is encoded, along with phonological information, within lexical representations.

### *In what way is lexical activation affected by indexical information?*

We used a set of different measures to evaluate exactly how our experimental manipulation of the voice-to-referent congruency affected lexical activation. Our main analyses were based on a curve-fitting procedure designed to evaluate the dynamic aspects of lexical activation. Specifically, we used a logistic equation to fit the curves of individual participants' looks to the target. This process allowed us to extract two sets of parameters per participant – one set for each of the two experimental conditions. Crucially, these parameters can be directly mapped onto psychologically meaningful aspects of lexical activation (e.g., speed of activation), thus allowing a fine-grained and meaningful comparison between experimental conditions.

Experiments 1 and 3 revealed a significant main effect of voice congruency on the slope parameter of the logistic function. The direction of this effect indicated a significantly faster build-up of target activation in the voice-congruent condition. In addition, all three experiments

revealed a significant main effect of voice congruency on the crossover parameter, indicating earlier activation of the target in the voice-congruent condition.

On the one hand, the two effects (on slope and crossover) can be thought of as theoretically codependent, given that a faster build-up of activation should lead to an earlier overall activation. According to this rationale, the more robust effect of voice congruency on the crossover parameter (compared to the slope) may simply reflect a discrepancy in our ability to capture meaningful differences in each of the two measures (that is, it may be more difficult to detect changes in speed of activation).

On the other hand, the two parameters can be interpreted independently, which means that the difference between the two effects may be theoretically meaningful. For example, the earlier crossover in the congruent compared to the incongruent trials could reflect a facilitation of the *onset* of lexical activation. This facilitation could be due to the system being faster at finding a match between the input and the stored representations. Once this match is achieved, lexical activation of the target can be initiated. Conversely, the less robust effect on the slope could be indicating that indexical information may still play a role at facilitating lexical activation even after a match has been achieved, but its contribution is not as critical as it is at the early stages of this process.

Although it will ultimately be interesting to tease apart these two possibilities, regardless of this resolution, the current study provides the first evidence that we know of that a talker's voice affects the *dynamics* of lexical activation and not just the overall degree of activation at a specific point in time.

### ***Indexical effects within and outside the lexicon***

As we outlined in the Introduction, indexical information may affect spoken word recognition directly (i.e., as part of lexical representations; see Figure 1.A) and/or indirectly (i.e., outside the mental lexicon; see Figure 1.B). Dissociating these two accounts in practice is extremely difficult, because they both predict an indexical effect. This is why we took special care to create a situation in which we would be able to detect evidence specifically for the former, direct kind of indexical effect, if indexical information is indeed included in lexical representations. We explain our rationale in the following paragraphs.



We begin by noting that indexical effects on lexical access can in some circumstances reflect voice information being utilized outside the mental lexicon. For example, as recently proposed by Cai et al. (2017), listeners can build speaker-models, which they then use to infer certain characteristics of the speaker (e.g., gender, age). They use these inferences during spoken word recognition to guide activation of the word's meaning (in Cai et al.'s study, this involved the disambiguation of semantically ambiguous words). Crucially, according to this account, voice information is not part of lexical representations, it is not stored in the mental lexicon, and it can only affect referent selection in an indirect way (i.e., via voice-driven inferences about the speaker). This conclusion was mainly based on their finding that when words were spoken in a neutral accent, participants used the accent of adjacently presented words to interpret them, suggesting that participants utilized voice information for a set of words presented together, rather than on a token-by-token basis.

The findings reported by Cai et al. (2017; see also Creel & Tumlin, 2011) provide evidence that listeners can use information about a talker's voice to access lexical meaning in an indirect way (i.e., outside the mental lexicon). This idea is also consistent with neuroimaging studies showing two partially dissociated functional pathways for processing linguistic content versus talker identity (Belin, Fecteau, & Bédard, 2004; Myers & Theodore, 2017). However, it is important to understand that the experimental paradigms used in these studies tested for inference-based use of indexical information. Logically, the fact that such an inference-based use exists says absolutely nothing about whether indexical information may *also* be encoded as part of lexical representations. In contrast, our experimental paradigm was specifically designed to prevent participants from using voice information in an explicit way, allowing us to look for indexical effects within the mental lexicon.

We have already described another possible way that voice information could bias referent selection outside the mental lexicon – via the direct association of voice information to specific images (see Supplementary Analyses). To assess this possibility, we analyzed trials in which the speaker's voice was associated with one or more of the filler items in the display. Even though we found no support for voice-picture associative effects, we conducted further analyses to test whether such associations might be the driving force behind our main effects. These analyses produced no support whatsoever for this possibility.

*Detecting effects of indexical information on semantic processing*

Previous studies have found very little evidence that indexical information is in direct contact with the lexical-semantic network. For example, Kittredge, Davis, and Blumstein (2006), as well as Lee and Zhang (2018), used a lexical decision task in which the target word (e.g., *queen*) was preceded by a semantic/associative prime (e.g., *king*) spoken by either the same or a different speaker. The semantic/associative priming effect was not modulated by speaker match in either of the two studies. Negative findings for lexical indexical effects were also reported by Luthra, Fox, and Blumstein (2018), who examined indexical effects on the false recognition of words. Participants studied sets of semantically related words (e.g., *butter, jam, crust*) and were then asked to recognize those items. During test, studied items were intermixed with semantically related novel items (e.g., *bread*). The results showed that false recognition of novel items was affected by speaker voice only when subjects actively attended to talker identity during encoding.

These results seem to challenge the view that indexical information can affect access to meaning in a direct and automatic manner and, in that respect, they seem to stand in contrast to our findings. However, in interpreting these null effects one must carefully consider the differences between studies in terms of the experimental paradigms, measures, and stimuli. For example, all three studies mentioned above used familiar words. This limits the kind of control an experimenter can have over the nature of lexical representations involved in a task and, critically, the indexical information these representations may already carry; if listeners have an established lexical representation for a given item, they may have learned that, at least for that item, indexical information is a dimension irrelevant to meaning. Similarly, adult listeners may have learned to generally ignore this kind of information because past experience has shown that it is not a relevant cue to spoken word recognition.

Crucially, this last point, that listeners may learn to ignore indexical information, is in line with the idea that they do not usually rely on this kind of information; however, it does not logically follow from this point that indexical information is not encoded within lexical representations. What we asked here is not what listeners *usually* do, but rather what they *can* do, because the answer to this question can provide a better understanding of the mechanisms that subserve lexical processing. That is, if the underlying mechanism is such that it allows for indexical

information to be encoded as part of lexical representations, then indexical effects should be observable when the conditions allow for it. In fact, as Pufahl and Samuel (2014) argued, the rapid pace of speech input may oblige listeners to store it without “cleaning up” potentially irrelevant information, with later memory processes effectively winnowing the available information to aspects that are common across multiple encounters (Goldinger, 1998).

In sum, a combination of conditions may need to be in place for indexical effects (intrinsic to the mental lexicon) to become detectable. Such conditions may include a) using novel lexical representations (with full control over the indexical information attached to them), b) preventing explicit use of voice information (which could overshadow any direct effects of indexical information on lexical access), and c) using a sensitive measure that can detect fine changes in the dynamics of lexical activation; all of which were present in our study.

### ***Implications for the mental lexicon***

Our results provide support for the hypothesis that voice information can be encoded within lexical representations. As such, they are consistent with the idea that the mental lexicon includes episodic traces rather than abstract units (Goldinger, 1996, 1998; Johnson, 2005, 2006; Palmeri et al., 1993), even though as Goldinger (1998) noted, a lexicon with episodic representations is quite capable of abstraction across these episodes, using basic memory properties.

The idea that the mental lexicon includes episodic traces does not imply that lexical representations are stored in episodic memory. In other words, the present work speaks to the content of the mental lexicon, rather than its location. Our findings are most consistent with distributed views of the mental lexicon, in which lexical representations can be conceptualized as carrying different kinds of information – including information that has been traditionally considered non-linguistic (Elman, 2009; Gaskell & Marslen-Wilson, 1997, 2002; Goldinger, 1998). For example, both the Distributed Cohort Model (DCM; Gaskell & Marslen-Wilson, 1997, 2002) and the “Echo” model (Goldinger, 1998) view words as multidimensional vectors that can contain different kinds of information (including talker’s voice information). Both the DCM and the Echo model can be seen as attempts to move the notion of the mental lexicon away from the traditional linguistic depiction of a list of word entries, with the entries including information about a word’s meaning and syntactic role. The motivation for the change is simple:

The traditional mental lexicon seems too limited to account for a growing body of evidence that lexical representations must be richer than the original conception.

Elman (2009) took this position to its logical extreme, and suggested that we should abandon the notion of a mental lexicon altogether. His analysis was driven by an investigation of the syntactic specificity that many verbs have. As he looked at how tightly a verb may control the form of its arguments, Elman concluded that a truly adequate lexical representation would need to include essentially all of the properties involved in event comprehension, with an event comprising consistently co-occurring pieces of information of various types. In his framework, a word is always processed in context, i.e., as part of an event, and acts as an external stimulus that alters the current state of the system by means of activating other information with which the given word/context combination has frequently co-occurred in the past. In this approach, a lexical representation is conceptualized as a set of similar patterns of activation trajectories. Elman suggested that since all of the event information was needed to understand verbs and their arguments, it was not clear why that information should be represented both as the experienced events, and as some kind of lexical duplication of this information. The idea of abandoning the concept of a mental lexicon altogether (in favor of an event-based description of language processing) is debatable; however, this approach highlights the difficulty in determining which types of information are encoded into lexical representations and the importance of context in word processing.

Elman's (2009) notion of event-based representation is very similar to the episodes that are represented in Goldinger's (1998) model. In both cases, experiences are stored in memory, and the language system relies on these stored experiences. Consistent with this view of the lexicon as being grounded in "episodes" or "events", Pufahl and Samuel (2014) demonstrated that listeners are better at recognizing words the second time they hear them, when a background sound is identical to a sound that was present the first time a given word was heard compared to when the background sound is different. Their experiments were modeled on classic indexical studies (e.g., Goldinger, 1996) showing that matching voice information facilitates word recognition. The new finding was that even apparently unrelated co-occurring information (e.g., the sound of a particular dog barking) has a similar effect. Just as Elman (2009) suggested, words are linked to a wealth of information beyond the well-defined entries in the traditional lexicon. These results, in line with the present work, provide strong evidence that during spoken word

recognition listeners encode different types of information that go beyond what has traditionally been considered linguistic.

It is this confluence of evidence for an embarrassment of riches in the lexicon that has driven the emergence of non-traditional views of the lexicon. Across a range of tasks, there is evidence for verb representations to need the most detailed, context-limited information (Elman, 2009), for voice information (Goldinger, 1996) and arbitrary background sounds (Pufahl & Samuel, 2014) to be stored in the lexicon, and for implicit use of voice information on referent selection found in the three experiments here. In all of these cases, the non-linguistic information has been shown to affect performance in word recognition tasks, both in terms of the ability to recognize words in noise, and in terms of the dynamic time course of recognition.

### *Lexical representations or episodic memory traces?*

Given this large and growing set of findings that call for a fundamental rethinking of what the lexicon looks like, is there a way to rescue the traditional lexicon? The most promising approach to doing so would be to attribute all of these “non-linguistic” effects to something other than lexical processing. According to this account, the mental lexicon can still be viewed as linguistically abstract, if we attribute the results of our three experiments to voice-image-word associations between non-linguistic memory traces stored in episodic memory. Indeed, perhaps with the exception of Elman's (2009) arguments about verbs, one might attribute the apparently-too-rich lexical findings in the literature to effects stemming from episodic memory, rather than from lexical access.

We see two problems with this alternative. First, there is a fundamental problem with invoking episodic memory: A defining property of episodic memory is its role in conscious recollection. In fact, this is the core principle that differentiates it from semantic memory (e.g., Tulving, 2002). In the current study we asked the participants in Experiment 1 if they had noticed anything about the role of the different talkers, and none of them was aware that there was any such role, let alone the linkage of a voice to a particular picture. The same lack of awareness was found in Pufahl and Samuel's (2014) study: Subjects were at chance on a post-test that called for judging whether a particular word-sound pair was one to which they had been previously exposed. Because the defining property of episodic memory is its being the source for conscious recollection, it is difficult to imagine how information in episodic memory is producing the

observed effects when participants have no conscious awareness of the relevant information. Perhaps this would be possible if episodic memory could host memory traces that do not meet the current definitional criteria of episodic memory. Since our study does not speak to this, further work would be needed to rule out this possibility.

Second, an explanation of the current results (and previous ones) that relies on episodic memory encounters a logical problem. Our task required participants to recognize the new words and access their meaning as they do with previously known words. Given this task, our effect could be attributed to episodic memory representations, in one of three ways. One possibility is that both the observed congruency effect and spoken word recognition take place within episodic memory, which would mean that episodic memory can support spoken word recognition. If episodic memory can explain the observed word recognition results, then it seems to need the same properties as the “real” lexicon (e.g., the similarity space among words). By extension, if both systems have similar properties, then it would be difficult to generate predictions that can distinguish the two accounts.

An alternative explanation is that while participants were recognizing words (using the mental lexicon), their eye-movements were also driven by episodic memory traces of voice-image-word associations. However, since participants were not aware of these associations, this explanation depends on the assumption that episodic memory can host memory traces that are not consciously accessible, and thus do not meet the current definitional criteria of episodic memory. As mentioned earlier, our study does not speak to this possibility and, thus, further work is needed to rule out this explanation.

Finally, an alternative explanation is to assume that participants were not recognizing words; they were doing some kind of nonlinguistic matching task. If mapping speech input to visual referents can be mistaken for spoken word recognition, then this leads to a similar difficulty in generating testable predictions that differ from the predictions of the proposed account.

In sum, we acknowledge that there are alternative interpretations of our results that may rely on episodic memory or a nonlinguistic matching process and further work is needed to examine these alternatives. That said, we believe that the collection of empirical and logical problems makes the alternatives less plausible than an explanation that instead treats lexical representations as multidimensional memory structures.

*Learning multidimensional lexical representations*

The proposed conception of the lexicon requires some account of how these richer lexical structures could develop. That is, how are different types of information encoded in lexical representations? Our results suggest that the answer to this question lies in the concept of *systematic co-occurrence*. This idea links the present work to a broad literature on associative learning, according to which lexical representations are the result of the systematic co-occurrence of different types of information (McMurray, Horst, Toscano, & Samuelson, 2009; Siskind, 1996; Smith & Yu, 2008; Yu & Smith, 2007). A dominant account within this literature is that proposed by Yu and Smith (2007, 2012), according to whom words can be learned via a cross-situational learning mechanism that tracks co-occurrence statistics between words and objects across different situations. For example, the word “table” may occur in the presence of a table, a plate, and a fork, but in a different situation it may occur in the presence of a table and a vase. By tracking co-occurrence statistics over time, it becomes clear that the object *table* is the referent of the word “table”, because it is the object that most frequently co-occurs with it. In this sense, word-learning can be viewed as another instance of associative learning (Regier, 2005). Our findings fit nicely into this type of associative learning framework, according to which the mental lexicon indiscriminately incorporates all kinds of information that systematically co-occur during the processing of a given word (see Pufahl & Samuel, 2014, for a similar view).

According to this view, the robustness of indexical effects should depend on the degree to which indexical information systematically co-occurs with the other pieces of information that comprise a word. For example, talker's voice effects should be modulated by the number of different voices associated with a given word. If this is true, indexical information should play a more important role during the early stages of word learning, when a word-form has only been associated with a small number of voices. At that stage, it may still be unclear whether voice information is a relevant cue for lexical access or not. As Pufahl and Samuel (2014) suggested, the time-critical nature of speech perception may oblige listeners to encode information without a great deal of decision-making about future utility of the information; storage and retrieval processes can effectively demote the impact of stored information that is not later useful. Then, as a listener is progressively exposed to more versions of the novel word, and as long as variability is not systematically linked to any aspect of lexical access, it should become clear that

indexical information is not a relevant cue (this would be compatible with the results reported by Brown and Gaskell, 2014; also see McMurray & Jongman, 2011, and Rost & McMurray, 2009, for a similar account of how talker variability can be harnessed to support listeners in learning which dimensions are relevant for phoneme categorization). However, even if indexical effects weaken as a result of increased voice variability, this should not be taken as evidence that indexical information is not encoded in lexical representations. Rather, as more and more related episodes are experienced (e.g., hearing a particular word spoken by various talkers, in various contexts), any particular indexical detail will only play a role in access if it is consistently present.

### ***Conclusion***

The present experiments demonstrate that listeners can encode indexical information as part of lexical representations and use it to access lexical meaning. These findings best fit into an episodic and distributed view of the mental lexicon, where words are conceptualized as multidimensional vectors comprised of both phonological and non-phonological (commonly referred to as extra-linguistic) information. There is of course ample evidence that humans operate using more abstract information, but such abstractions can be derived in real time from the accumulation of episodic information encoded in multidimensional vectors. Ultimately, a person's experiences in the world are inherently episodic, and any abstractions must be derived from these.

### **Acknowledgements**

Support for this project was provided by the Spanish Ministry of Science and Innovation, Grant #PSI2014-53277 and # PSI2017-82563-P awarded to A.G.S., the Spanish Ministry of Economy and Competitiveness, Juan de la Cierva-Formación fellowship awarded to E.C.K., and the Spanish Ministry of Economy and Competitiveness, "Severo Ochoa" Programme for Centres/Units of Excellence in R&D (SEV-2015-490). We would also like to thank James McQueen for his suggestions about this paper, and an anonymous reviewer for suggesting the last of the supplementary analyses.



## References

- Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: evidence for continuous mapping models. *Journal of Memory and Language*, 38(4), 419–439.
- Apfelbaum, K., Blumstein, S., & McMurray, B. (2011). Semantic priming is affected by real-time phonological competition: evidence for continuous cascading systems. *Psychonomic Bulletin & Review*, 18(1), 141–9.
- Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: neural correlates of voice perception. *Trends in Cognitive Sciences*, 8(3), 129–135.
- Borovsky, A., Elman, J. L., & Kutas, M. (2012). Once is enough: n400 indexes semantic integration of novel word meanings from a single exposure in context. *Language Learning and Development : The Official Journal of the Society for Language Development*, 8(3), 278–302.
- Borovsky, A., Kutas, M., & Elman, J. (2010). Learning to use words: event-related potentials index single-shot contextual word learning. *Cognition*, 116(2), 289–96.
- Bowers, J. S. (2000). In defense of abstractionist theories of repetition priming and word identification. *Psychonomic Bulletin & Review*, 7(1), 83–99.
- Bradlow, A. R., Nygaard, L., & Pisoni, D. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception & Psychophysics*, 61(2), 206–219.
- Brown, H., & Gaskell, M. G. (2014). The time-course of talker-specificity and lexical competition effects during word learning. *Language, Cognition and Neuroscience*, 29(9), 1163–1179.
- Cai, Z. G., Gilbert, R. A., Davis, M. H., Gaskell, M. G., Farrar, L., Adler, S., & Rodd, J. M. (2017). Accent modulates access to word meaning: evidence for a speaker-model account of spoken word recognition. *Cognitive Psychology*, 98, 73–101.
- Cooper, A., & 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).

- Coutanche, M., & Thompson-Schill, S. (2014). Fast mapping rapidly integrates information into existing memory networks. *Journal of Experimental Psychology: General*, *143*(6), 2296.
- Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2008). Heeding the voice of experience: the role of talker variation in lexical access. *Cognition*, *106*(2), 633–64.
- Creel, S. C., & Tumlin, M. (2011). On-line acoustic and semantic interpretation of talker information. *Journal of Memory and Language*, *65*(3), 264–285.
- Cutler, A., & Weber, A. (2007). Listening experience and phonetic-to-lexical mapping in 12. In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the 16th International Congress of Phonetic Sciences* (pp. 43–48). Dudweiler: Pirrot.
- Davis, M., & Gaskell, M. G. (2009). A complementary systems account of word learning: neural and behavioural evidence. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *364*(1536), 3773–800.
- Dumay, N., & Gaskell, M. G. (2007). Sleep-associated changes in the mental representation of spoken words. *Psychological Science*, *18*(1), 35–39.
- Dumay, N., & Gaskell, M. G. (2012). Overnight lexical consolidation revealed by speech segmentation. *Cognition*, *123*(1), 119–32.
- Duñabeitia, J. A., Crepaldi, D., Meyer, A. S., New, B., Pliatsikas, C., Smolka, E., & Brysbaert, M. (2017). MultiPic: a standardized set of 750 drawings with norms for six european languages. *The Quarterly Journal of Experimental Psychology*, 1–24.
- Elman, J. L. (2009). On the meaning of words and dinosaur bones: lexical knowledge without a lexicon. *Cognitive Science*, *33*(4), 547–582.
- Farris-Trimble, A., & McMurray, B. (2013). Test–Retest reliability of eye tracking in the visual world paradigm for the study of real-time spoken word recognition. *Journal of Speech, Language, and Hearing Research*, *56*(4), 1328–1345.
- Fernandes, T., Kolinsky, R., & Ventura, P. (2009). The metamorphosis of the statistical segmentation output: lexicalization during artificial language learning. *Cognition*.
- Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: a distributed model of speech perception. *Language and Cognitive Processes*, *12*(5–6), 613–656.

- Gaskell, M. G., & Marslen-Wilson, W. D. (1999). Ambiguity, competition, and blending in spoken word recognition. *Cognitive Science*, 23(4), 439–462.
- Gaskell, M. G., & Marslen-Wilson, W. D. (2002). Representation and competition in the perception of spoken words. *Cognitive Psychology*, 45(2), 220–266.
- 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.
- Goldinger, S. D. (1998). Echoes of echoes? an episodic theory of lexical access. *Psychological Review*, 105(2), 251–279.
- Goldinger, S. D. (2007). A complementary-systems approach to abstract and episodic speech perception. In *16th International Congress of Phonetic Sciences* (pp. 49–54). Saarbrücken, Germany.
- González, J., & McLennan, C. T. (2007). Hemispheric differences in indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology. Human Perception and Performance*, 33(2), 410–24.
- Johnson, K. (2005). Speaker normalization in speech perception. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 363–389). Malden, MA: Blackwell Publishing Ltd.
- Johnson, K. (2006). Resonance in an exemplar-based lexicon: the emergence of social identity and phonology. *Journal of Phonetics*, 34(4), 485–499.
- Kapnoula, E. C., & McMurray, B. (2016). Newly learned word-forms are abstract and integrated immediately after acquisition. *Psychonomic Bulletin and Review*, 23(2), 491–499.
- Kapnoula, E. C., Packard, S., Gupta, P., & McMurray, B. (2015). Immediate lexical integration of novel word forms. *Cognition*, 134, 85–99.
- King, E., & Sumner, M. (2015). Voice-specific effects in semantic association. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

- Lee, C.-Y., & Zhang, Y. (2018). Processing lexical and speaker information in repetition and semantic/associative priming. *Journal of Psycholinguistic Research*, *47*(1), 65–78.
- Lindsay, S., & Gaskell, M. G. (2013). Lexical integration of novel words without sleep. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *39*(2), 608–622.
- Luce, P. A., & Lyons, E. A. (1998). Specificity of memory representations for spoken words. *Memory & Cognition*, *26*(4), 708–715.
- Luthra, S., Fox, N. P., & Blumstein, S. E. (2018). Speaker information affects false recognition of unstudied lexical-semantic associates. *Attention, Perception, & Psychophysics*, 1–19.
- McClelland, J. L., & Elman, J. L. (1986). The trace model of speech perception. *Cognitive Psychology*, *18*(1), 1–86.
- 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), 419–57.
- McLennan, C., & 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), 306–21.
- McMurray, B., Horst, J., Toscano, J., & Samuelson, L. (2009). Towards an integration of connectionist learning and dynamical systems processing: case studies in speech and lexical development. In J. Spencer, M. Thomas, & J. McClelland (Eds.), *Towards an Integration of Connectionist Learning and Dynamical Systems Processing: Case Studies in Speech and Lexical Development*. London, UK: Oxford University Press.
- McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, *118*(2), 219–46.
- McMurray, B., Samelson, V. M., Lee, S. H., & Tomblin, J. B. (2010). Individual differences in online spoken word recognition : implications for sli. *Cognitive Psychology*, *60*(1), 1–39.
- McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, *86*(2), B33–B42.

- McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, *30*(6), 1113–1126.
- Myers, E. B., & Theodore, R. M. (2017). Voice-sensitive brain networks encode talker-specific phonetic detail. *Brain and Language*, *165*, 33–44.
- Norris, D. (1994). Shortlist: a connectionist model of continuous speech recognition. *Cognition*, *52*(3), 189–234.
- Oleson, J., Cavanaugh, J., McMurray, B., & Brown, G. (2015). Detecting time-specific differences between temporal nonlinear curves: analyzing data from the visual world paradigm. *Statistical Methods in Medical Research*, 0962280215607411.
- Palmeri, T. J. T., Goldinger, S. D., & Pisoni, D. B. D. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(2), 309–328.
- Pufahl, A., & Samuel, A. G. (2014). How lexical is the lexicon? evidence for integrated auditory memory representations. *Cognitive Psychology*, *70*, 1–30.
- Regier, T. (2005). The emergence of words: attentional learning in form and meaning. *Cognitive Science*, *29*(6), 819–865.
- Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, *12*(2), 339–349.
- Salverda, A. P., & Tanenhaus, M. K. (2017). The visual world paradigm. In A. M. B. de Groot & P. Hagoort (Eds.), *Research Methods in Psycholinguistics*. Hoboken, NJ: Wiley.
- Samuel, A. G. (2011). Speech perception. *Annual Review of Psychology*, *62*(1), 49–72.
- Seedorff, M., Oleson, J., Brown, G., Cavanaugh, J., & McMurray, B. (2017). Bdots: bootstrapped differences of time series. r package version 0.1.15.
- Seedorff, M., Oleson, J., Cavanaugh, J., & McMurray, B. (2017). Eyetracking analysis in r. *R Package Version 0.1.15*.
- Seedorff, M., Oleson, J., & McMurray, B. (2018). Detecting when timeseries differ: using the bootstrapped differences of timeseries (bdots) to analyze visual world paradigm data (and more). *Journal of Memory and Language*, *102*, 55–67.

- Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*(1–2), 39–91.
- Smith, L., & Yu, C. (2008). *Infants rapidly learn word-referent mappings via cross-situational statistics*. *Cognition* (Vol. 106).
- Sumner, M., Kim, S. K., King, E., & McGowan, K. B. (2014). The socially weighted encoding of spoken words: a dual-route approach to speech perception. *Frontiers in Psychology*, *4*, 1015.
- Tamminen, J., Payne, J. D., Stickgold, R., Wamsley, E. J., & Gaskell, M. G. (2010). Sleep spindle activity is associated with the integration of new memories and existing knowledge. *The Journal of Neuroscience*, *30*(43), 14356–60.
- Van Berkum, J. J. A., van den Brink, D., Tesink, C. M. J. Y., Kos, M., & Hagoort, P. (2008). The neural integration of speaker and message. *Journal of Cognitive Neuroscience*, *20*(4), 580–591.
- Xie, X., Earle, F. S., & Myers, E. B. (2018). Sleep facilitates generalisation of accent adaptation to a new talker. *Language, Cognition and Neuroscience*, *33*(2), 196–210.
- Yu, C., & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*.
- Yu, C., & Smith, L. B. (2012). Modeling cross-situational word-referent learning: prior questions. *Psychological Review*, *119*(1), 21–39.