Audiovisual semantic interactions between linguistic and non-linguistic stimuli: The time-courses and categorical specificity

Yi-Chuan Chen[1,2,*] & Charles Spence[1]

[1]Department of Experimental Psychology, University of Oxford, Oxford, UK

[2]Department of Medicine, Mackay Medical College, New Taipei City, Taiwan

Running title: Crossmodal Semantic Congruency

Word counts: 11587 words

Re-submitted to: *Journal of Experimental Psychology: Human Perception and Performance*

Resubmitted: 19 Feb 2018

*Corresponding:
Dr. Yi-Chuan Chen
Department of Medicine, Mackay Medical College
No.46, Sec.3, Zhongzheng Rd., Sanzhi Dist., New Taipei City, 252, Taiwan
Tel: +886 2 26360303 #1255
Fax: +886 2 26361295
Email: ycchen@mmc.edu.tw

Abstract

We examined the time-courses and categorical specificity of the crossmodal semantic congruency effects elicited by naturalistic sounds and spoken words on the processing of visual pictures (Experiment 1) and printed words (Experiment 2). Auditory cues were presented at seven different stimulus onset asynchronies (SOAs) with respect to the visual targets, and participants made speeded categorization judgments (living vs. non-living). Three common effects were observed across two experiments: Both naturalistic sounds and spoken words induced a slowly-emerging congruency effect when leading by 250 ms or more in the congruent as compared to the incongruent condition, and a rapidly-emerging inhibitory effect when leading by 250 ms or less in the incongruent condition as opposed to the noise condition. Only spoken words that did not match the visual targets elicited an additional inhibitory effect when leading by 100 ms or when presented simultaneously. As compared to non-linguistic stimuli, the crossmodal congruency effects associated with linguistic stimuli occurred over a wider range of SOAs, and occurred at a more specific level of the category hierarchy (i.e., the basic level) than required by the task. A comprehensive framework is proposed to provide a dynamic view regarding how meaning is extracted during the processing of visual or auditory linguistic and non-linguistic stimuli, therefore contributing to our understanding of multisensory semantic processing in humans.

Public Significance Statements

Contrasting with the intuition that hearing a dog barking, or someone saying the word "DOG", will always help us to perceive the dog or read the word "DOG", we demonstrate that such facilitation only occurred when a congruent auditory cue was presented 250 ms or more before seeing the visual picture/word. We further demonstrate that a spoken word (for example, "DOG"), as compared to a naturalistic sound (e.g., a dog barking), influenced visual processing over a longer temporal interval (up to 1 second). In addition, the spoken word "DOG" seemed to provide more

specific semantic information (i.e., dog-like things) than the sound of a dog barking (i.e., animal-like things). The current study therefore demonstrates the conditions in which meaningful auditory stimuli modulate the processing of visual pictures and words. This is critical for designing the most efficient and effective multisensory warning or entertaining systems with modern technologies.

Keywords: Crossmodal; Meaning; Picture; Word; Sound; Lexical; Categorization

Introduction

When walking along the street on a foggy dawn, for example, hearing a dog's bark or else hearing someone say the word "dog" may well help you recognize an ambiguous far away creature as a dog. Naturalistic sounds (i.e., the dog's bark) and spoken words (i.e., the word "DOG") correspond to non-linguistic and linguistic auditory stimuli that convey meaning to the listener. They are different, however, in terms of their source and processing. A naturalistic sound is produced by the associated object itself, such as a barking sound and a dog, or a piano sonata and a piano (e.g., Ballas & Howard, 1987). On the other hand, spoken words and the associated visual objects originate from different sources: The former are produced by humans in order to refer to the latter. Neuroimaging studies demonstrate that the semantic processing of naturalistic sounds and spoken words is mainly distributed in common brain areas but lateralised to one or the other side. Specifically, naturalistic sounds are often processed predominantly in the right superior temporal areas, whereas spoken words, in the left superior temporal areas (Dick et al., 2007; Hocking & Price, 2009; Plante, van Petten, & Senkfor, 2000; Thierry, Giraud, & Price, 2003; van Petten & Rheinfelder, 1995; though see Cummings et al., 2006).

The goal of the present study was to compare crossmodal semantic congruency effects elicited by naturalistic sounds and spoken words on the processing of visual pictures and printed words – the latter two corresponding to non-linguistic and linguistic visual stimuli, respectively. Understanding the semantic modulation of visual stimuli after the presentation of auditory cues is interesting for the following reasons: Vision typically plays a dominant role in multisensory perception, especially when it comes to detecting or identifying objects (e.g., Colavita, 1974; Koppen, Alsius, & Spence, 2008; see Spence, Parise, & Chen, 2011, for a review). Additionally, less time is required for semantic access for visual than for auditory stimuli (e.g., Kim, Porter, & Goolkasian, 2014; Weatherford, Mills, Porter, & Goolkasian, 2015). Specifically, the meaning of a visual picture can potentially be accessed around 100 ms after its onset (see Fabre-Thorpe, 2011; Potter, 2014, for reviews), but the meaning of a naturalistic sound is accessed later at around 150 ms

after onset (see Murray & Spierer, 2009, for a review). Yuval-Greenberg and Deouell (2009) reported a larger semantic modulation from visual picture to naturalistic sound than the other way round when the two stimuli were presented simultaneously. Such asymmetrical effect can be easily explained by the faster semantic access for a visual picture than a naturalistic sound, so the meaning of the former is more likely to modulate the processing of the latter than vice versa. We therefore consider that it is more important to try and understand whether, and under what conditions, the semantic congruency of auditory information modulates visual perception than the reverse condition in humans.

A cognitive model of interactions between linguistic and non-linguistic stimuli

Almost three decades ago, Glaser and Glaser (1989) proposed a comprehensive model designed to try and explain the Stroop-like interference that they observed between visual pictures and printed words (see Figure 1). According to their model, there are two separate cognitive systems: a semantic system where the concept and knowledge regarding objects is stored, and a lexical system that contains lexical knowledge without any semantic capability. Glaser and Glaser suggest that a signature of stimulus interactions in the semantic system is the *semantic gradient effect* (e.g., Fox, Shor, & Steinman, 1971; Klein, 1964). That is, the magnitude of the congruency effect is positively correlated with the semantic relatedness between the cue and target (e.g., *cat* and *rabbit* are closer than *cat* and *table*). On the other hand, the absence of the semantic gradient effect suggests that cue-target interactions probably occurred in the lexical system.

Insert Figure 1 about here

According to Glaser and Glaser's (1989) model, after perceptual processing, visual pictures access their meaning directly, whereas printed words activate the corresponding lexical representations first, and the latter serve as the mediators by which access is granted to the associated semantic representations (see also Endress & Potter, 2012). The different processing

routes for visual pictures and words, then, account for their different time-courses of semantic access as demonstrated by subsequent event-related potential (ERP) studies: Visual pictures in the different superordinate categories (e.g., animal vs. non-animal)[1] elicit discrepant activities from 75 ms after picture onset (Thorpe, Fize, & Marlot, 1996; VanRullen & Thorpe, 2001). However, when reading printed words, the brain activities by 150 ms after word onset are associated with orthographic processing (Hauk, Davis, Ford, Pulvermüller, & Marslen-Wilson, 2006; Holcomb & Grainger, 2006), and brain activities associated with their meaning occur later (i.e., the N400, a negative activity at 200-600 ms after stimulus onset, see Kutas & Federmeier, 2011; Kutas & Hillyard, 1980).

According to Glaser and Glaser's (1989) model, the processing of spoken words resembles printed words in that the semantic representations are accessed through the associated lexical representations. The auditory lexical representation is activated when the acoustic information of a spoken word is sufficient to be discriminated from other phonologically-similar words, and this time-point (called the discrepancy point) typically occurs before the word's full presentation (e.g., Marslen-Wilson, 1987). An ERP study by van Petten, Soulson, Rubin, Plante, and Parks (1999) demonstrates that the N400 component associated with a word's meaning starts at the discrepancy point of a given word, even when the word happens to be presented in a congruent syntax provided by the preceding words.

The time-course of crossmodal semantic congruency effects

Chen and Spence (2011) extended Glaser and Glaser's (1989) model by suggesting that naturalistic sounds, similar to visual pictures, access the associated meaning directly after auditory sensory processing. This claim was based on the results of those studies demonstrating that

[1] Concrete objects belonging to a category share common attributes. In taxonomy, categories are hierarchically-structured. Supposedly, there is a most inclusive level at which all of the members of a category share most common attributes meanwhile sharing least common attributes with members of other categories. This level is known as the "basic level", such as *dog*, *car*, and *shirt* (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). With respect to the basic-level categories, the superordinate categories are more abstract and the members share less common features, such as *mammal*, *vehicle*, and *clothing*. In contrast, the subordinate categories are even more specific than basic level categories, and the members in different subordinate categories share common attributes, such as *Labrador* and *Beagle*, *sports car* and *sedan*, and *sweatshirt* and *polo shirt*.

naturalistic sounds access their associated meaning more rapidly than spoken words (Cummings et al., 2006; Saygin, Dick, & Bates, 2005). Consistent with this notion, Chen and Spence (2011, 2017) reported a series of experiments demonstrating that only naturalistic sounds, rather than spoken words, elicit a semantic congruency effect on visual sensitivity (*d'*) in a picture detection task when the auditory cue leads by 350 ms[2]. In addition, the crossmodal interference in a visual picture categorization task elicited by incongruent naturalistic sounds occurs at a shorter leading interval (240 ms) than in the case elicited by incongruent spoken words (399 ms; Chen & Spence, 2013).

When the stimulus onset asynchrony (SOA) was prolonged to 1000 ms, however, only spoken words, rather than naturalistic sounds, elicit a semantic congruency effect on the *d'* in the picture detection task (Chen & Spence, 2017). This result is consistent with the advantage for spoken words over naturalistic sounds at such long SOAs using a picture verification task that has been reported in previous studies. Specifically, the participants' reaction time (RT) was shorter when matching a picture to a spoken word than to a naturalistic sound (Boutonnet & Lupyan, 2015; Edmiston & Lupyan, 2015; Lupyan & Thompson-Schill, 2012).

The above-mentioned results suggest that the optimal SOAs for naturalistic sounds and spoken words to prime visual pictures differ. The first goal of the present study was therefore to verify the time-courses of audiovisual semantic congruency effects elicited by naturalistic sounds and spoken words on *visual picture* processing. In addition, the current study was further extended to test the time-courses of the modulations by naturalistic sounds and spoken words on *printed word* processing, which have only been compared previously at an SOA of 0 ms (Iordanescu, Grabowecky, & Suzuki, 2011; Iordanescu, Guzman-Martinez, Grabowecky, & Suzuki, 2008). By testing semantic interactions between stimuli in terms of the 2 (naturalistic sound/spoken word) x 2 (visual picture/printed word) design, the hope was that it would lead to a better understanding of the

[2] Nevertheless, it should be noted that the interaction between spoken words and visual pictures can be speeded-up so as to be comparable with that between naturalistic sounds and visual pictures by varying the task demands. Specifically, when participants had to detect as well as identify the picture by reporting its name, both types of auditory stimuli primed the pictures at the 350 ms SOA (see Chen & Spence, 2011).

time required for each type of stimulus to access its respective meaning and interact crossmodally, as happens commonly in daily life (see also Noppeney, Josephs, Hocking, Price, & Friston, 2008).

The category specificity of crossmodal semantic congruency effects

In terms of the conventional category hierarchy (i.e., the superordinate, basic, and subordinate levels, see Rosch et al., 1976), it has long been suggested that the basic level is the "entry point" to semantic access concerning objects (e.g., Grill-Spector & Kanwisher, 2005; Jolicoeur, Gluck, & Kosslyn, 1984). However, accumulating evidence suggests a coarse-to-fine semantic processing of visual pictures (e.g., Mack & Palmeri, 2015; see Clarke, 2015; Fabre-Thorpe, 2011; for reviews). Specifically, objects belonging to different superordinate-level categories can be discriminated at around 110 ms after stimulus onset, while those of different basic-level categories are discriminated somewhat later (Clarke, Devereux, Randall, & Tyler, 2015; Wu, Crouzet, Thorpe, & Fabre-Thorpe, 2014). In a similar vein, the coarse-to-fine semantic processing of naturalistic sounds has also been demonstrated (Murray, Camen, Andino, Bovet, & Clarke, 2006, Murray, Camen, Spierer, & Clarke, 2008; see Murray & Spierer, 2009, for a review).

The second goal of the present study was to examine the categorical specificity of audiovisual interactions between linguistic and non-linguistic stimuli that are associated with a basic-level object (e.g., a dog image, its barking, and the word "DOG"). We manipulated three levels of semantic relatedness between the auditory cue and the visual target for each combination of linguistic and non-linguistic stimuli. That is, congruent cues matched the targets at the basic level (e.g., dog); mismatched cues belonged to either the same category (related condition) or different categories (incongruent condition) with respect to the targets at the superordinate level of the category hierarchy (i.e., as need to make a living vs. non-living distinction).

Overview of the current study

We compared the audiovisual semantic congruency effects elicited by naturalistic sounds or spoken words on the processing of visual pictures (Experiment 1) and printed words (Experiment 2). A speeded categorization task in which the participants had to judge (using manual responses)

whether the visual target belongs to the class of living or non-living things was used. The categorization task has proved useful when it comes to probing human semantic processing, because the participant's decision is based on the output from the semantic system rather than the lexical selection for articulation (e.g., Mädebach, Wöhner, Kieseler, & Jescheniak, 2017; Roelofs, 1992). Three levels of semantic relatedness between the auditory cue and the visual target were manipulated (congruent, related, and incongruent). A neutral condition in which white noise was presented, as well as a no-sound condition (where only the visual target was presented) were both included for comparison. Seven SOAs from auditory cue leading by 1000 ms to lagging by 250 ms with respect to the visual target were used in order to demonstrate the time-courses of the various audiovisual semantic interactions. The current study extends Chen and Spence's (2013) study in the following critical regards: First, auditory semantic modulations on visual picture and printed words were compared, while only the former was tested in Chen and Spence (2013). Second, the range of SOAs between the auditory cue and visual target was extended in the current study relative to Chen and Spence (2013) in which SOAs within ±399 ms were tested. Finally, the related condition was included in the current study in addition to the congruent and incongruent conditions that were tested in Chen and Spence (2013).

Experiment 1: Picture categorization task

In Experiment 1, the participants had to judge whether the object presented as an outline drawing was a living thing or not. A task-irrelevant auditory cue, either a naturalistic sound or a spoken word, was presented at one of seven SOAs in order to probe the audiovisual interactions at the different stages of human information processing.

Methods

Participants

Thirty volunteers (10 males, mean age of 22.4 years) took part in this study in exchange for course credits or 10 pounds (UK Sterling). The participants were native English speakers or

bilinguals who had started to learn English by the age of six years. All of the participants had normal or corrected-to-normal vision and normal hearing by self-report, and all were naïve as to the purpose of the study. Written informed consent was obtained from each of the participants before the study began. The study was approved by the Medical Sciences Inter Divisional Research Ethics Committee, University of Oxford (MSD-IDREC-C1-2014-143).

The number of participants was determined before conducting the study in order to achieve adequate statistical power. An analysis using G*Power (version 3.1.9.2; Faul, Erdfelder, Lang, & Buchner, 2007) suggests that testing 30 participants ($\alpha = 0.05$, power = 0.95), the 3-way interaction (Sound Type x SOA x Congruency) that we were mainly interested in would reach $\eta_p^2 = 0.036$ (the effect size *f* provided by G*Power was transformed to η_p^2 based on the Appendix of Lakens', 2013, paper). There was no previous study using the same experimental design, so we compared the number of participants and effect size reported in the closest relevant studies. Specifically, in Chen and Spence's (2013) study, 14 participants were tested and a significant 3-way interaction (Sound Type x SOA x Congruency) was observed ($\eta_p^2 = 0.164$, 90% confidence interval (CI) = [0.030, 0.191], see Lakens, 2014; Smithson, 2001; Wuensch, 2016). In Ostarek and Huettig's (2017) Experiment 2, 33 participants were tested and a significant 2-way interaction (SOA x Congruency) was observed ($\eta_p^2 = 0.103$, 90% CI = [0.012, 0.184]). It therefore seemed likely that testing 30 participants would lead to an effect size falling within the range of 90% CI reported in previous studies.

Apparatus and stimuli

The visual stimuli were presented on a 23-inch LED monitor (60 Hz refresh rate) controlled by a personal computer. The participants sat at a viewing distance of around 58 cm from the monitor in a dimly-lit chamber. Twenty-four outline-drawings (12 living and 12 non-living things) taken from Snodgrass and Vanderwart (1980) and Bates et al. (2003) were used as visual targets. The original pictures and their mirror images were used. All of the target pictures covered an area of less than 5.5°×5.5° and were presented in the center of a white background.

The auditory stimuli (8 bit mono; 22,500 Hz digitization) were those used in Chen and Spence's (2017) study (see Appendix A). All of the naturalistic sounds were downloaded from www.findsounds.com (on 08/08/2008). Only one sound was used as an exemplar for each object. The spoken words were the most commonly agreed name used to refer to a given picture in the studies of Snodgrass and Vanderwart (1980) and Bates et al. (2003). The spoken words were produced by a female native English speaker. The naturalistic sound and the spoken word that refer to the same object were edited to have the same duration of 350 ms for the 14 one-syllable words, 450 ms for the seven two-syllable words, and 500 ms for the three- (or more) syllable words (three words). White noise presented for 350 ms was chosen as the neutral auditory stimulus in order to control for different levels of alerting or distracting effect elicited by the presentation of an auditory stimulus at each SOA (e.g., Bertelson & Tisseyre, 1969; Robertson, Mattingley, Rorden, & Driver, 1998). The sound pressure levels (SPL, in terms of the root mean square values) of all the auditory stimuli were equalized. The auditory stimuli were presented over closed-ear headphones and their peak ranged in loudness from 31-51 dB SPL.

Design

Three factors: Sound Type (naturalistic sound vs. spoken word), SOA (-1000, -500, -250, -100, 0, 100, and 250 ms), and Congruency (congruent, related, incongruent, or noise), were manipulated. The factors of Sound type and SOA were blocked (e.g., Chen & Spence, 2013; Glaser & Glaser, 1989; Roelofs, 2005), giving rise to 14 blocks of trials in total. Previous research demonstrates that similar results were observed when the SOA was either manipulated on the basis of blocks or mixed within a block (Donohue, Appelbaum, Park, Roberts, & Woldorff, 2013; Roelofs, 2010). The seven SOA blocks containing naturalistic sounds or spoken words were grouped into the first or second experimental sessions. The order in which the naturalistic sound and spoken word sessions were conducted was counterbalanced across participants.

The SOAs were chosen based on previous research. The four negative SOAs indicate those trials in which the sound was presented first. The -1000 ms SOA (corresponding to 500 to 650 ms

inter-stimulus interval, ISI) is the interval in which the advantage of spoken words over naturalistic sounds has been demonstrated (Chen & Spence, 2017; Edmiston & Lupyan, 2015; Lupyan & Thompson-Schill, 2012). In the -500 ms SOA condition, the presentation of the auditory cue had been completed by the onset of the target picture. In the -250 ms SOA condition, the target picture onset while the auditory cue was still being presented. These two SOAs are close to the -350 ms SOA that was used to demonstrate a crossmodal semantic priming effect elicited only by naturalistic sounds rather than by spoken words (Chen & Spence, 2011, 2017). The -100 and 0 ms SOAs have commonly been used to demonstrate crossmodal Stroop-like interference (e.g., Meyer & Schriefers, 1991; Roelofs, 2005; Shimada, 1990; Yuval-Greenberg & Deouell, 2009). The two positive SOAs (100 and 250 ms) indicate those trials in which the auditory cue was presented after the onset of the picture. These two SOAs were included because previous research has demonstrated that the presentation of meaningful auditory stimuli can still modulate participants' performance in response to visual targets even when the auditory stimuli briefly lagged visual targets (e.g., Chen & Spence, 2010, 2013; Roelofs, 2005; Shimada, 1990). The order of presentation of the seven SOA blocks was randomized.

The semantic congruency factor was mixed within blocks of trials. In the congruent trials, the auditory cue matched the target picture (e.g., a barking sound and a dog picture). In the related trials, the auditory cue did not match the target picture, though they both belonged to the same living or non-living category (e.g., a cow's mooing sound and a picture of a dog). In the incongruent trials, the paired sounds and pictures belonged to different living vs. non-living categories (e.g., a guitar chord and a picture of a dog). The same visual and auditory stimuli were used in each condition, and related and incongruent cue-target pairs were the same across different SOAs and sound types. Hence, any differences in participants' performance could simply be attributed to the variation of the three manipulated factors, rather than to any difference of stimulus identities (i.e., no item variability across conditions in the current design).White noise was presented in the noise condition. In each SOA block, all 24 pictures were presented four times, accompanied by either a congruent,

related, or incongruent cue, or else by the noise; that is, the auditory cue was congruent with the target picture only in a quarter of the trials. The order of presentation of these 96 trials was randomized within a block of trials.

Procedure

In the main experimental session, the participants initiated a block of trials by pressing the "ENTER" key on the keyboard in front of them. In each trial, a blank frame was presented for 1411 ms, followed by a target picture which was presented until the participants responded or 2000 ms had elapsed (see Figure 2). The participants had to decide whether they had just seen a picture of a living thing by pressing the "RIGHT ARROW" key or else a picture of a non-living thing by pressing the "LEFT ARROW" key. In each condition, participants' accuracy and RT were obtained from equal numbers of living- and non-living-thing trials. The participants were informed that they should respond to the picture rather than to the sound, and that to respond as rapidly as possible, while maintaining a high degree of accuracy was more important. They were not informed about the relationships between sounds and pictures. At the end of each block of trials, the participants would see their overall accuracy in that block, and they were asked to maintain their accuracy at above 90%.

Insert Figure 2 about here

Each participant completed a no-sound block (i.e., the target picture was presented alone) containing all 24 target pictures used in the main experiment first in order to familiarize him/herself with the task. Next, a practice session consisting of 16 trials (including the four congruency conditions) was conducted prior to the main experiment. The SOA in this practice block was 0 ms and the sound type was the same as the participant encountered first during the experiment proper. The stimuli in the practice session were not used in the main experiment. The experimenter made sure that the participant followed the task instructions before proceeding onto the main experiment.

After the participant had completed the first session (including the seven SOA blocks containing either naturalistic sounds or spoken words), the no-sound block was again tested in order to estimate the baseline RTs for the unisensory picture categorization task. Finally, the participant completed the second session containing the seven SOA blocks of the other sound type. The entire experiment lasted for approximately one hour.

Predictions

A significant difference between participants' performance in the congruent and incongruent conditions provides a measure of semantic congruency effect. The semantic congruency effect can be decomposed into the facilitation in the congruent condition, and/or inhibition in the incongruent condition, when comparing each condition to the neutral condition respectively.

In addition to the congruent and incongruent condition, adding the related condition would provide a further measure of the underlying mechanism of cue-target interactions (see Table 1, for examples of RT patterns). The cue-target interaction in the semantic system would be expected to yield a shorter RT in the congruent than in the related condition at the basic level of the category hierarchy; nevertheless, congruent and related cues would be expected to induce a similar RT at the superordinate level since both provide the same categorical information (i.e., living or non-living). In both cases, shorter RTs in the related than in the incongruent condition should be observed given the weaker semantic relatedness in the latter case[3]. The cue-target interaction in the lexical system would be expected to yield shorter RTs in the congruent than in the related and incongruent conditions (i.e., RT: Congruent < Related = Incongruent), because the prime and target are only matched in the first condition.

The last possibility is that cue and target may interact at the response level. Specifically, the congruent and related cues may induce a compatible response to the target because they all belong to the same response category. However, incongruent cues may induce an incompatible response to

[3] In Glaser and Glaser's (1989) Experiment 6, they demonstrated that the semantic gradient effects differed in different tasks. Specifically, in the picture categorization task, a smaller inhibitory effect was induced by the presentation of a related picture than by an incongruent picture. On the other hand, in the picture naming task, the reversed pattern was observed (i.e., larger inhibitory effect by a related than an incongruent picture). Given that the categorization task was used in the current study, we therefore follow the results in the former case.

the target because they belong to different response categories. This cue-target compatibility at the response level would be expected to yield a similar RT in the congruent and related conditions, both are shorter than in the incongruent condition (i.e., RT: Congruent = Related < Incongruent). Hence, the accounts of response compatibility and semantic congruency at the superordinate level predict similar RT patterns. In order to tease apart these two mechanisms, therefore, one would need to compare the results between different cue-target combinations. If a type of auditory cue reliably induced such a pattern of results, then the response compatibility would provide the most parsimonious explanation; otherwise, the account of semantic congruency at the superordinate level would provide a better explanation for a particular cue-target combination.

Insert Table 1 about here

Results

The accuracy data were submitted to a three-way analysis of variance (ANOVA) with the factors of Sound Type, Congruency, and SOA (see Table 2). Overall accuracy reached 97.7 % (*SE* = 0.2%), and none of the main effects or interactions was significant (all Fs $\leq$ 1.42, ps $\geq$ .24, η_p^2s $\leq$ 0.05). We therefore focus on the analyses of the RT data.

Insert Table 2 about here

The median RT (for correct response trials only) in each condition for each participant was used in order to reduce any potential influence of positive skewed distributions in the RT data (Chen & Spence, 2013; Hays, 1973)[4]. The median RT data were submitted to a three-way ANOVA.

[4] Preliminary analyses of the RT data were conducted. The RT data in each condition were fitted with an ex-Gaussian distribution (Lacouture & Cousineau, 2008). The results demonstrated that the width in the long-RT side was 2.4 times wider than in the short-RT side, suggesting a positive skewed distribution. Nevertheless, using either inverse- or log-transformed RT data to normalize the data may lead to the distortion of additive/interactive effects (see Lo & Andrews, 2015). We have also compared the results in terms of mean RT and median RT, and they demonstrated a similar pattern

The main effect of Congruency was significant ($F(3,87) = 27.91$, $MSE = 559.01$, $p < .001$, $\eta_p^2 = 0.49$), as was every level of interactions associated with this factor (Congruency x Sound Type: $F(3,87) = 6.10$, $MSE = 460.37$, $p < .005$, $\eta_p^2 = 0.17$; Congruency x SOA: $F(18,522) = 7.15$, $MSE = 480.94$, $p < .001$, $\eta_p^2 = 0.20$; Congruency x Sound Type x SOA: $F(18,522) = 1.83$, $MSE = 427.51$, $p < .05$, $\eta_p^2 = 0.06$). No other main effect or interaction was significant (Fs ≤ 1.61, ps $\geq .15$, η_p^2s ≤ 0.05).

A couple of two-way follow-up ANOVAs on the factors of Congruency and SOA were then conducted for the naturalistic sounds and spoken words, separately. For naturalistic sounds (see Figure 3A), the main effect of Congruency was significant ($F(3,87) = 8.40$, $MSE = 340.37$, $p < .001$, $\eta_p^2 = 0.23$). Post-hoc tests (pairwise t-tests with Holm-Bonferroni correction) demonstrate that the RT was shorter in the congruent (486 ms) and noise condition (488 ms) than in the incongruent condition (495 ms, both $t(29) \geq 3.69$, ps $< .01$). The interaction between Congruency and SOA was significant ($F(18,522) = 1.89$, $MSE = 452.10$, $p < .05$, $\eta_p^2 = 0.06$). The simple main effect of Congruency (see Table 3A) was significant at the -500 and -250 ms SOAs (Fs ≥ 5.04, ps $< .005$, η_p^2s ≥ 0.14). Post-hoc tests demonstrated a significant semantic congruency effect in terms of the shorter RT in the congruent and related condition as compared to the incongruent condition at the -500 ms SOA (both $t(29) \geq 3.27$, ps $< .05$). A significant semantic congruency effect was also observed between the congruent and incongruent condition at the -250 ms SOA ($t(29) = 3.60$, $p < .01$), mainly attributable to an inhibitory effect in the incongruent as compared to the noise condition ($t(29) = 2.92$, $p < .05$).

For spoken words (see Figure 3B), the main effect of Congruency ($F(3,87) = 22.90$, $MSE = 679.01$, $p < .001$, $\eta_p^2 = 0.44$) was significant. Post-hoc tests demonstrate that the RT was shorter in the congruent (484 ms) than in the related (496 ms) and noise (495 ms) conditions, and the RTs in these three conditions were shorter than in the incongruent condition (505 ms, all $t(29) \geq 3.62$, ps

as a function of SOA. However, the SE of the mean RT was larger than the SE of the median RT (11.6 vs. 10.2 in Experiment 1, and 14.0 vs. 11.9 in Experiment 2). Taken together, we used the median RT to represent the individual participant's performance in each condition in order to reduce the influence of outliners and positive skewed RT in each condition. The median RT also provides an estimate of participants' performance with smaller between-participant variabilities than the mean RT.

< .005). The interaction between Congruency and SOA was significant ($F(18,522) = 7.37$, $MSE = 456.36$, $p < .001$, $\eta_p^2 = 0.20$). The simple main effect of Congruency (see Table 3B) was significant at the -1000, -500, -250, -100, and 0 ms SOAs (Fs ≥ 3.78, ps $< .05$, η_p^2s ≥ 0.11). In summary, at the -1000 and -500 ms SOAs, the RT was shorter in the congruent than in the other three conditions (all $t(29) \geq 3.77$, ps $< .005$). The RT was shorter in the related condition than in the noise condition at the -1000 ms SOA, and it was shorter in the related condition than in the incongruent condition at the -500 ms SOA (both $t(29) \geq 3.39$, ps $< .01$). At the -250 ms SOA, an inhibitory effect was documented in the incongruent condition as compared to the other three conditions (all $t(29) \geq 3.48$, ps $< .01$). Finally, at the -100 and 0 ms SOAs, an inhibitory effect was documented in the related and incongruent conditions as compared to the noise condition (all $t(29) \geq 2.77$, ps $< .05$).

The median RT of picture categorization in the no-sound block without any auditory cue was 465 ms ($SE = 10.5$). This is the shortest RT as compared to any of the conditions in which an auditory cue was presented in Experiment 1 (the second shortest was 467 ms when a congruent spoken word was leading by 1000 ms). This result suggests that the presentation of the auditory cue, if it had any influence, delayed the participant's response, suggesting a higher loading of information processing in the multisensory than unisensory visual condition (see also Chen & Spence, 2013).

Insert Figure 3 and Table 3 about here

Discussion

The presentation of an auditory cue consisting of either a naturalistic sound or spoken word elicited two similar effects on participants' picture categorization performance, though at slightly different SOAs: The first was a slow effect of semantic congruency (i.e., the RT was shorter in the congruent than in the incongruent condition) when the auditory cue led the target picture by 500 ms or more (i.e., -1000 and -500 ms SOAs). Second, a rapid effect of semantic congruency, mainly

attributable to an inhibition in the incongruent condition as compared to the noise condition, was observed when the auditory cue led by 250 ms (i.e., at the -250 ms SOA). There was an additional effect that was only elicited by the spoken words: An inhibitory effect was observed when a related or incongruent spoken word was presented 100 ms before, or at the same time as, the onset of the picture (i.e., -100 and 0 ms SOAs). No effect was observed when the auditory cue was presented after the onset of the picture (i.e., 100 and 250 ms SOAs).

The slow semantic congruency effect was observed across a narrower range of SOAs for naturalistic sounds (only at the -500 ms SOA) than for spoken words (at the -1000 and -500 ms SOAs). In addition, the congruency effect was less pronounced for naturalistic sounds given that none of the congruent, related, or incongruent conditions was significantly different from the noise condition. Most critically, similar facilitatory effects were observed in the congruent and related conditions as compared to the incongruent condition, suggesting that the congruency effect elicited by naturalistic sounds was based on the living/non-living distinction. Such crossmodal semantic interactions can occur at two possible stages of processing, either the semantic interactions at the superordinate level, or as a result of response compatibility (see Table 1). Experiment 2, where the crossmodal semantic congruency effect elicited by naturalistic sounds on printed words was tested, would provide further evidence to verify either one of the accounts.

The slow semantic congruency effect elicited by spoken words was observed at the -1000 ms SOA and was still there at the -500 ms SOA. This result is unequivocally attributed to the facilitation in the congruent condition as compared to the noise condition. Critically, at the -500 ms SOA, a semantic gradient effect was observed such that the RT increased from the congruent to the related and then the incongruent condition. This result clearly indicates that spoken words and visual pictures interact in the semantic rather than the lexical system, and specifically at the basic level of the category hierarchy. This result also rules out the response compatibility account. In summary, as compared to naturalistic sounds, spoken words primed the visual picture across larger leading intervals and at the more specific categorical level.

The rapid semantic congruency effect induced by naturalistic sounds at the -250 ms SOA was mainly attributed to the interference by the presentation of an incongruent auditory cue as compared to the noise control. This result suggests that 250 ms provided sufficient time for naturalistic sounds to access their meaning (which takes around 156-215 ms, see Murray et al., 2008), and the incongruent meaning then interfered with the processing of target pictures. This interference effect is not a result of response compatibility, because the RTs were not significantly different in the related condition as compared to the incongruent condition.

The rapid interference by incongruent spoken words as compared to the noise was at the -250, -100, and 0 ms SOAs[5]. A similar inhibitory effect was induced by the related cue as compared to the noise at the -100 and 0 ms. These two effects should be considered together: At the -250 ms SOA, the longer RT in the incongruent than in the related condition (i.e., a semantic gradient effect) suggests that this effect occurred in the semantic system rather than in the lexical system (Glaser & Glaser, 1989). At the -100 and 0 ms SOAs, the similar interference effects in the related and incongruent conditions (i.e., no semantic gradient effect) resembles the Stroop-like interference induced by either printed or spoken words in the picture categorization task (Glaser & Düngelhoff, 1984; Stuart & Carrasco, 1993). This interference was specific to spoken words without a semantic gradient effect, suggesting that the lexical representations should have been involved (Glaser & Glaser, 1989). Taken together, spoken words influence the participants' picture categorization judgments through the connection from the lexical to the semantic representations, and this

[5] The time required to access lexical representations, as well as the associated semantic representations, of spoken words are positively correlated with word length (Marslen-Wilson, 1987; van Petten et al., 1999). The -250 ms SOA is the critical condition because around 70% of acoustic signals of one-syllable words had been presented, whereas only 50-56% of the acoustic signals associated with the words with multiple syllables were presented. We therefore analyzed the semantic congruency effects elicited by the one-syllable versus multisyllabic words at the -250 ms SOA separately. The two-way ANOVA with the factors of Congruency (congruent, related, and incongruent) and Length (one or multiple syllables) were conducted. The results demonstrated a significant main effect of Congruency ($F(2,58) = 15.52$, $MSE = 1132.14$, $p < .001$, $\eta_p^2 = 0.35$) and their interaction ($F(2,58) = 8.72$, $MSE = 866.31$, $p < .001$, $\eta_p^2 = 0.23$). The one-way follow-up ANOVA demonstrated that, for one-syllable words, the Congruency effect was significant ($F(2,58) = 26.90$, $MSE = 833.96$, $p < .001$, $\eta_p^2 = 0.48$). Post-hoc tests demonstrated that the RT was shorter in the congruent (470 ms) and related (479 ms) than in the incongruent condition (523 ms, both $t(29) \geq 5.32$, ps $< .001$). By contrast, for multisyllabic words, the Congruency effect was not significant ($F(2,58) = 1.20$, $p = .31$). These results therefore suggest that the semantic congruency effects at the -250 ms SOA were susceptible to word length. This SOA is perhaps too short for multisyllabic words to access their meaning. Given the naturalistic sounds and spoken words were matched in terms of their durations, trials with a naturalistic sound cue can be separated and analyzed in the same way. However, this did not reveal a Congruency x Length interaction ($F(2,58) = 1.12$, $p = .33$), thus suggesting that the Congruency effects were similar at the -250 ms SOA irrespective of the length of the naturalistic sounds.

processing was demonstrated at the range of -250 to 0 ms SOAs. When the spoken word led by only 100 or 0 ms, the interference came from the mismatched lexical representations; when leading by 250 ms, the spoken word should have accesses its semantic representation, giving rise to the semantic gradient effect.

In summary, three effects were observed in Experiment 1: The first was a slow semantic congruency effect that occurred at different SOAs for naturalistic sounds (at -500 ms) and spoken words (at -1000 and -500 ms). Critically, the semantic interactions between spoken words and visual pictures occurred at the basic level, which is more specific than the interactions between naturalistic sounds and visual pictures at the superordinate level. The second effect was the rapid semantic congruency effect at the -250 ms SOA induced by both naturalistic sounds and spoken words. This effect was attributed to the interference of semantically incongruent information. Finally, only the related and incongruent cue which was a spoken word elicited another interference effect at the -100 and 0 ms SOAs. This effect suggests that the spoken words automatically accessed the lexical code and thence disturbed the processing of the visual pictures via lexical-semantic interactions.

Experiment 2: Word categorization task

In Experiment 2, the participants had to categorize a printed word. An auditory cue, either a naturalistic sound or a spoken word, was presented at one of the seven possible SOAs.

Methods

A new group of 30 volunteers (nine males, mean age of 24.1 years) from the same participant pool took part in this study. The only difference from Experiment 1 was that the visual target was an English word printed in black and presented in the centre of a white background on the monitor. The word was capitalized for the first letter and printed in Arial font ranging from 1.38° x 0.69° to 4.73° x 0.69° (width x height) depending on the length of the word. The participants had to judge whether the word refers to a living thing or not as accurately and rapidly as possible. Three factors:

Sound Type (naturalistic sound vs. spoken word), SOA (-1000, -500, -250, -100, 0, 100, and 250 ms), and Congruency (congruent, related, incongruent, or noise), were manipulated. The other details were the same as in Experiment 1.

Results

The accuracy data were submitted to a three-way ANOVA with the factors of Sound Type, SOA, and Congruency (see Table 4). The overall accuracy reached 97.2 % (*SE* = 0.3%). The main effect of Congruency was significant ($F(3,87) = 4.03$, $MSE = 0.001$, $p < .05$, $\eta_p^2 = 0.12$). Post-hoc tests revealed that the accuracy was higher in the congruent condition (97.5%) than in the incongruent condition (96.8%; $t(29) = 3.23$, $p < .05$). The two-way interaction between Congruency and SOA was significant ($F(18,522) = 3.00$, $MSE = 0.001$, $p < .001$, $\eta_p^2 = 0.09$). The simple main effect of Congruency was significant at the -500, -100, and 0 ms SOAs (all Fs ≥ 3.67, ps $< .05$). Post-hoc tests demonstrated a higher accuracy in the congruent, related, and noise conditions than in the incongruent condition at the -500 ms SOA (all $t(29) \geq 3.04$, ps $< .05$), and a higher accuracy in the congruent and related conditions than the incongruent condition at the -100 ms SOA ($t(29) \geq 2.84$, ps $< .05$). None of the post-hoc tests reached statistical significance at the 0 ms SOA (all $t(29) \leq 2.77$, ps $\geq .06$). In summary, the accuracy in the incongruent condition was lower than some of the other conditions at the -500 and -100 ms SOAs. However, it should be noted that the RT was longest in the incongruent condition at these two SOAs (see Figures 4A and 4B), thus suggesting that there was no speed-accuracy trade-off.

Insert Table 4 about here

The median RT data were submitted to a three-way ANOVA with the factors of Sound Type, Congruency, and SOA. Two main effects were significant: Congruency ($F(3,87) = 31.61$, $MSE = 1646.68$, $p < .001$, $\eta_p^2 = 0.52$) and SOA ($F(6,174) = 2.50$, $MSE = 7840.83$, $p < .05$, $\eta_p^2 = 0.08$). There was a significant two-way interaction between Congruency and SOA ($F(18,522) = 6.02$, *MSE*

= 910.68, $p < .001$, $\eta_p^2 = 0.17$). Critically, the three-way interaction was significant ($F(18,522) = 1.73$, $MSE = 868.36$, $p < .05$, $\eta_p^2 = 0.06$).

A couple of two-way ANOVAs were then conducted on the Congruency and SOA factors for the naturalistic sounds and spoken words, separately. For naturalistic sounds (see Figure 4A), the main effect of Congruency was significant ($F(3,87) = 22.45$, $MSE = 702.14$, $p < .001$, $\eta_p^2 = 0.44$). Post-hoc tests revealed that the RT was shorter in the congruent (555 ms) than in the related condition (566 ms), and the RT in the congruent, related, and noise (559 ms) conditions were all shorter than in the incongruent condition (574 ms; all $t(29) \geq 3.52$, ps $< .01$). The Congruency x SOA interaction was significant ($F(18,522) = 2.45$, $MSE = 758.64$, $p < .005$, $\eta_p^2 = 0.08$). The simple main effect of Congruency (see Table 5A) was significant at the -500, -250, -100, and 0 ms SOAs (Fs ≥ 3.00, ps $< .05$, η_p^2s ≥ 0.09). Post-hoc tests demonstrated a facilitatory effect in terms of the RT being shorter in the congruent than in the other three conditions at the -500 and -250 ms SOAs (all $t(29) \geq 3.54$, ps $< .01$), and an inhibitory effect in the incongruent as compared to the noise condition at the -250, -100, and 0 ms SOAs (all $t(29) \geq 2.95$, ps $< .05$).

When the auditory cues were spoken words (see Figure 4B), both main effects of Congruency ($F(3,87) = 18.80$, $MSE = 2094.22$, $p < .001$, $\eta_p^2 = 0.39$) and SOA ($F(6,174) = 2.90$, $MSE = 6597.40$, $p < .05$, $\eta_p^2 = 0.09$) were significant. Post-hoc tests demonstrate that the RT was shorter in the congruent (552 ms) than in the related (570 ms) and noise conditions (565 ms), and they were all shorter than in the incongruent condition (585 ms, all $t(29) \geq 2.35$, ps $< .05$). As compared to the -1000 ms SOA, the RT in the 0 ms SOA was significantly longer ($t(29) = 3.27$, $p < .05$). The Congruency by SOA interaction was significant ($F(18,522) = 5.03$, $MSE = 1020.39$, $p < .001$, $\eta_p^2 = 0.15$). The simple main effect of Congruency (see Table 5B) was significant at all seven SOAs (Fs ≥ 3.16, ps $< .05$, η_p^2s ≥ 0.09). At the -1000, -500, and -250 ms SOAs, a facilitatory effect was observed such that the RT was shorter in the congruent than in the other three conditions (all $t(29) \geq 2.59$, ps $< .05$). In addition, the RT was shorter in the related than the incongruent condition at the -250 ms SOA ($t(29) = 3.76$, $p < .01$). At the -100 and 0 ms SOA, an inhibitory effect was observed

such that the RT was longer in the incongruent than in the congruent and noise conditions (all $t(29) \geq 2.99$, ps $< .05$); this effect lasted until the 100 ms SOA when compared to the noise condition ($t(29) = 4.26$, $p < .01$). Across this time window, the RT in the related condition was longer than the congruent condition at the -100 ms SOA ($t(29) = 2.75$, $p < .05$), and longer than the noise condition at the 0 ms SOA ($t(29) = 4.09$, $p < .01$).

Insert Figure 4 and Table 5 about here

RTs for the printed word categorization in the no-sound block without an auditory cue was 533 ms (*SE* = 9.7), which was longer than the RT in the following three conditions: a congruent naturalistic sound presented at the -250 ms SOA (532 ms), a congruent spoken word presented at the -1000 and -500 ms SOA (525 and 521 ms, respectively). However, the presentation of a congruent auditory cue was not able to facilitate the participants' RT at a level that was statistically significant when compared to the RT in the no-sound block (all $t(29) \leq 1.35$, ps $> .18$; see also Chen & Spence, 2013).

Finally, we compared the participants' baseline performance in categorizing the visual pictures (Experiment 1) and printed words (Experiment 2) in the no-sound block. The accuracy was similar in the picture and word conditions (96.7% vs. 97.1%, $t(58) = 0.48$, $p = .63$, two-tailed). However, RTs were shorter for pictures than for words (465 ms vs. 533 ms, $t(58) = 4.73$, $p < .001$), thus suggesting a slower process of semantic access for printed words than for pictures (see also Glaser & Düngelhoff, 1984).

Discussion

Consistent with the results of visual picture categorization that were reported in Experiment 1, both types of auditory cues induced two similar semantic congruency effects on the categorization of printed words: the slowly-emerging semantic congruency effect and the rapidly-emerging inhibitory effect following the presentation of the incongruent cue. However, both of the effects

spread to later SOAs (see Table 6). It is possible, given that the time of semantic access was longer for printed words than visual pictures, that an auditory cue would be able to induce a similar effect even when it was presented later (i.e., at more positive SOAs). Note that none of the above effects were consistent with the pattern predicted by the response compatibility account. Finally, only spoken words in the related condition elicited a unique inhibitory effect as compared to the congruent condition at the -100 ms SOA and to the noise condition at the 0 ms SOA.

The slow semantic congruency effect elicited by naturalistic sounds on printed words (at the -500 and -250 ms SOAs) demonstrated a facilitatory effect which was only significant in the congruent condition, rather than in the related condition. This is different from the effect elicited by the same naturalistic sounds on visual pictures that both congruent and related sounds elicited similar congruency effects (see Experiment 1). Taken together, these semantic congruency effects were the outcomes of cue-target interaction, rather than simply depending on the semantic information provided by naturalistic sounds. Specifically, naturalistic sounds and visual pictures interact at the superordinate level, whereas naturalistic sounds and printed words interact at the basic level of the category hierarchy.

The slow semantic congruency effect elicited by spoken words on printed words (from -1000 ms to -250 ms SOAs) came from the shorter RT in the congruent than in the related, incongruent, and noise conditions, which should occur at the basic level of the category hierarchy. In addition, at the -250 ms, a significant semantic gradient effect (i.e., RT: Congruent < Related < Incongruent) was observed, which is similar to the effect by spoken words on visual pictures reported at the -500 ms SOA in Experiment 1. These results suggest that the related spoken words elicited a weaker and shorter-lived congruency effect than the congruent spoken words. Combining the results induced by naturalistic sounds and spoken words, it would appear that meaningful auditory cues and *printed words* only interact at the basic level of the category hierarchy, even when the demand of the word categorization task was set at the superordinate level.

The second effect, the rapid interference caused by the presentation of incongruent naturalistic sounds on printed words (from -250 to 0 ms SOA)[6], again had a longer time-course than on visual pictures (only observed at -250 ms SOA). By contrast, the rapid interference effect elicited by the incongruent spoken words on printed words occurred from the SOAs of -100 to 100 ms, which generally shifted to later SOAs as compared to the same effect on visual pictures (-250 to 0 ms SOAs). In addition, over the period from -100 to 0 ms SOAs, the RT in the related condition was similar to that reported in the incongruent condition, and both were longer than in the congruent and/or noise conditions. Such a pattern of results is associated with the interactions with the lexical representations given that no semantic gradient effect was observed (Glaser & Glaser, 1989). Taken together, spoken words interfered with the processing of printed words at SOAs from -100 to 100 ms, starting when they interact in the lexical system.

General Discussion

In the present study, we compared the time-courses of crossmodal semantic congruency effects elicited by naturalistic sounds or spoken words on visual pictures and printed words. When using a visual categorization task in terms of the living vs. non-living distinction, both types of auditory cues induced two similar effects at different time-courses: A slow semantic congruency effect elicited by the congruent as compared to the incongruent cue when leading by 250 ms or more, and a rapid inhibitory effect elicited by the incongruent cue as compared to the noise when leading by 250 ms or less. In addition, the related and incongruent spoken words elicited a unique inhibitory effect on both types of visual targets when leading by 100 ms or presented simultaneously. Most

[6] Once again, the analysis of separating one- vs. multiple-syllable words was conducted at the -250 ms SOA, and similar results to those reported in Experiment 1 were observed. For spoken words, the results of the two-way ANOVA with the factors of Congruency (congruent, related, and incongruent) and Length (one or multiple syllables) demonstrated a significant main effect of Congruency ($F(2,58) = 16.35$, $MSE = 956.73$, $p < .001$, $\eta_p^2 = 0.36$) and their interaction ($F(2,58) = 5.28$, $MSE = 1090.79$, $p < .01$, $\eta_p^2 = 0.15$). The Congruency effect was significant for one-syllable words ($F(2,58) = 27.08$, $MSE = 731.10$, $p < .001$, $\eta_p^2 = 0.48$). Post-hoc tests demonstrated that the RT was shorter in the congruent (479 ms) and related (485 ms) conditions than in the incongruent condition (526 ms, both $t(29) \geq 6.09$, $ps < .001$). By contrast, for multisyllabic words, there was no Congruency effect ($F(2,58) = 1.22$, $p = .30$). These results again suggest that the semantic congruency effects at the -250 ms SOA were susceptible to word length. The same analysis for naturalistic sounds demonstrated that the interaction for Congruency and Length was not significant ($F(2,58) = 1.02$, $p = .37$), thus suggesting that the congruency effect at -250 ms SOA was not susceptible to the length of the naturalistic sounds.

critically, the precise time-courses and the levels of category hierarchy where these effects occurred were different for each type of cue-target combination.

Since the publication of Glaser and Glaser's (1982) study examining the time-courses of the visual picture-word interactions, it has been repeatedly reported the dissociation of the slow semantic congruency effect by congruent cues and the rapid inhibitory effect resulting from incongruent cues in unisensory visual studies (see also Glaser & Düngelhoff, 1984; Glaser & Glaser, 1989; Roelofs, 2010), as well as in audiovisual studies (Chen & Spence, 2010, 2013; Donohue et al., 2013; Roelofs, 2005). Such a result is hard to explain if one only considers the time required for semantic access of the auditory cue, because this account predicts that both facilitatory and inhibitory effects would occur over the same time-course (e.g., the repetition priming and its counterpart, antipriming, see Marsolek, 2008). The response compatibility account *cannot* provide the most parsimonious explanation for these results either. Instead, the dissociable time-courses of the slow semantic congruency effect and the rapid inhibitory effect should be the outcomes of cue-target interactions at different stages of information processing.

The slowly-emerging semantic congruency effect

Both naturalistic sounds and spoken words enhanced the speed of categorization of visual pictures or printed words when they were congruent as compared to when they were incongruent. Three features of this effect are summarized as follows (see Table 6). First, the starting time of this congruency effect was associated with the type of auditory cue: Naturalistic sounds primed the visual targets when leading by 500 ms, and spoken words extended to an even longer interval when leading by 1000 ms (see also Boutonnet & Lupyan, 2015; Chen & Spence, 2017). The fact that the crossmodal semantic congruency effect persisted over a longer interval for spoken words than for naturalistic sounds should partly be attributed to the benefit for spoken words being briefly maintained in the phonological loop in the working memory system (see Baddeley, 2012, for a review). By contrast, naturalistic sounds are maintained either by being transferred into lexical codes and stored in the phonological loop, or else by using the auditory imagery capability (see

Snyder & Gregg, 2011; Soemer & Saito, 2015), and both processes takes extra time or cognitive resources[7].

The second feature of the slow semantic congruency effect is that its end is associated with the type of visual target. Specifically, when the target was a visual picture, this effect ended when the auditory cue led by 500 ms; in contrast, when the target was a printed word, this effect ended when the auditory cue led by 250 ms. This result can easily be explained in terms of the longer time required for semantic access of the printed words (see our results in the no-sound blocks; see also Glaser & Düngelhoff, 1984; Glaser & Glaser, 1989). That is, the longer-lasting semantic processing of visual targets leads to the later dissipation of this congruency effect.

The third feature of the slow semantic congruency effect is that different cue-target combinations demonstrated semantic interactions at different levels of the category hierarchy, even though all of the stimuli are associated with objects belonging to basic-level category. Specifically, naturalistic sounds primed visual pictures at the superordinate level in terms of the living vs. non-living distinction (RT: Congruent = Related < Incongruent). This result can be explained by the coarse-to-fine semantic processing of both naturalistic sounds (Murray et al., 2006, 2008) and visual pictures (Clarke et al., 2015; Mack & Palmeri, 2015; Wu et al., 2014), leading to their initial semantic interactions at the superordinate level. Note also that the semantic information at the superordinate level is sufficient to fulfill the current categorization task, while accessing the information at the basic-level is not necessary.

By contrast, when either one, or both, of the cue and target were linguistic stimuli, the largest priming effect was observed in the congruent condition that the cue and target were matched at the basic level of the category hierarchy (RT: Congruent < Related ≤ Incongruent). Such result was

[7] Using the picture categorization task, both Kim et al. (2014) and Weatherford et al. (2015) demonstrated that a congruent naturalistic sound elicited a facilitatory effect as compared to the same-category, different-category, and neutral conditions (equivalent to the related, incongruent, and noise conditions in the present study, respectively) only when the sound led by 1750 ms, rather than by 1250 ms. This effect resembles our results demonstrating crossmodal semantic congruency effect by a spoken word on the target picture at the -1000 ms SOA (i.e., a semantic interaction at the basic level). Hence, the semantic congruency effect elicited by a naturalistic sound over a longer interval (i.e., more than 1 sec) probably mediated by its accessing lexical code (i.e., the name of the object producing the sound) that is better maintained in the phonological loop. In turn, this lexical representation recurrently activates the semantic representation to prime the visual picture.

observed when the auditory cue was a spoken word in Experiment 1, and when the visual target was a printed word in Experiment 2. Following the model proposed by Glaser and Glaser (1989), both spoken and printed words access the semantic representations associated with their lexical representations, and so the lexical representation would be expected to activate the semantic representation at the basic level precisely (and perhaps spread to the superordinate and/or subordinate level afterwards).

Figure 5 represents the above crossmodal semantic interactions between linguistic and non-linguistic stimuli presented in vision or audition. The framework primarily adopts Glaser and Glaser's (1989) model in that two cognitive systems, the semantic and lexical, are separate but closely and bi-directionally connected. This connection is represented by a two-sided arrow between the "DOG" in the lexical system and the "DOG" in the semantic system (note that the connection looks long mainly because we would like to clearly demonstrate the structure of the two systems). The semantic access of non-linguistic stimuli (visual pictures and naturalistic sounds) is a hierarchical coarse-to-fine processing in the perceptual/semantic system. The semantic access of linguistic stimuli (printed and spoken words), instead, depends on the connection between lexical and semantic system at the level of semantic hierarchy to which the word corresponds.

Insert Figure 5 and Table 6 about here

Taken together, these three features suggest that the slow semantic congruency effect is an outcome of semantic interactions taking place between auditory cue and visual target. The leading auditory cue mainly facilitated the processing of the congruent visual target, rather than inhibiting the incongruent target. A neuroimaging study of crossmodal semantic priming consistently demonstrated that the audiovisual-association areas modulated visual category-specific areas mainly in the congruent rather than in the incongruent condition when auditory cues led visual pictures by 1300 ms (Adam & Noppeney, 2010). It would seem that, when the leading auditory cue has

sufficient time to access its meaning, it would facilitate the processing of the subsequently-presented visual target which is semantically congruent in a top-down manner.

In the current study, no crossmodal facilitation following the simultaneous presentation (i.e., at the 0 ms SOAs) of a congruent cue was observed, nor in previous studies using picture detection and picture categorization tasks (Chen & Spence, 2011, 2013). By contrast, some previous studies demonstrate that a simultaneous auditory cue facilitated a visual target when they are congruent in picture identification, identity discrimination, and visual search tasks (Chen & Spence, 2010; Iordanescu et al., 2008; Suied, Bonneel, & Viaud-Delmon, 2009; Yuval-Greenberg & Deouell, 2009). We suggest that task demand is a critical factor. When the response threshold of the visual target is rapidly achieved, such as in the picture detection task, the simultaneously-presented sound may have no time to access its meaning before completing the task. In contrast, when more detailed information of the visual target is accumulating for identification, the simultaneously-presented sound would have time to access its meaning and thence modulate visual processing (see Chen & Spence, 2013).

The rapid inhibitory effect elicited by incongruent and related cues

The presentation of an auditory cue that is semantically incongruent with the visual target reliably induced an inhibitory effect when it led by 250 ms or less. The time-course was associated with the type of both auditory cue and visual target: When the auditory cue was a naturalistic sound, the inhibitory effect occurred at the -250 ms SOA for both types of visual target, but extended to the 0 ms SOA when the target was a printed word. This result can be explained by the slower semantic access for printed words than for visual pictures, leading to a later ending of the effect for printed words.

When the auditory cue was a spoken word, the inhibitory effect by the incongruent cue was at the -250 to 0 ms SOAs for visual pictures, and shifted later to the -100 to 100 ms SOAs for printed words. At the -100 and 0 ms SOAs, similar interference induced by the related and incongruent cues suggests that this effect arose from the mismatched lexical codes (Glaser & Glaser, 1989). Taken

together, in the course of spoken word processing from the lexical to the semantic system, the processing of visual pictures was modulated by either lexical or semantic information depending on their SOAs, whereas spoken and printed words started to interact at the stage of lexical access.

This rapid inhibitory effect suggests that the incongruent auditory cue interfered with the perceptual processing of the visual target. Consistent evidence comes from a previous ERP study: When the auditory and visual stimuli were presented simultaneously, semantic modulations on visual processing occurred at around 150 ms after onset (Molholm, Ritter, Javitt, & Foxe, 2004). A neuroimaging study further demonstrates that the incongruent sound impaired the patterned activities associated with the visual objects in early visual cortices (V2 and V3, de Haas, Schwarzkopf, Urner, & Rees, 2013). Such auditory semantic interference on early visual processing can be considered as a form of crossmodal prediction error in the predictive coding scheme. That is, the brain uses information in one sensory modality to predict or modulate the processing of the incoming signals in another sensory modality (see Arnal, Wyart, & Giraud, 2011).

Implications of the results to human perceptual and cognitive processing

In contrast to the early view that perceptual and cognitive processing are distinct modules (Fodor, 1975; Pylyshyn, 1984), researchers have proposed alternative models for the human semantic system that are grounded in the perceptual systems to certain extents (see Barsalou, 2016; Lambon Ralph, Jefferies, Patterson, & Rogers, 2017; Mahon & Hickok, 2016, for recent reviews). For example, while retaining the core amodal semantic system, the processing in the perceptual, motor, and affective systems seems essential in providing detailed and context-dependent features for an enriched semantic representation (e.g., Binder & Desai, 2011; Mahon & Caramazza, 2008; Patterson, Nestor, & Rogers, 2007). Other models have proposed that the semantic system consists of a hierarchical convergence with each modality-specific system linking to the amodal system, thus giving rise to a "hub-and-spoke" structure (e.g., Damasio, 1989; Plaut, 2002; Pobric, Jefferies, & Lambon Ralph, 2010; Reilly, Peelle, Garcia, & Crutch, 2016). Finally, the models of grounded cognition exclude the amodal semantic system; instead, they suggest that the human semantic

systems are represented by the systems for perception (including vision, audition, and touch, etc.), action, and internal states that are used to process a given stimulus. Nevertheless, the grounded view may agree that there are convergent zones which are essentially *multisensory* (Barsalou, 1999, 2008; Gallese & Lakoff, 2005). We suggest that research on the topic of audiovisual semantic congruency provides critical insights regarding how crossmodal information converges and how its meaning is processed during human information processing (see Figure 5; see Barsalou, 2016).

Convergence zone of audiovisual semantic congruency

Consistent with all of the above models that suggest a convergence zone receiving multisensory information, psychologists and neuroscientists have tried to search for the core areas responsible for the processing of audiovisual semantic congruency. These areas include the superior temporal sulci, inferior temporal cortex, and perirhinal cortex (see Simanova, Hagoort, Oostenveld, & van Gerven, 2014; Taylor, Moss, Stamatakis, & Tyler, 2006; Werner & Noppeney, 2010; see Doehrmann & Naumer, 2008; Lewis, 2010, for reviews). Such evidence suggests that audiovisual integration/interactions involving semantic congruency occur in higher cortical areas than modality-specific sensory areas, consistent with the conventional view based on anatomical and functional structure of the human brain (Felleman & Van Essen, 1991).

A growing body of evidence nevertheless demonstrates extensive interactions between traditionally-defined areas that are modality-specific in the human cerebral cortex (see Driver & Noesselt, 2008; Ghazanfar & Schroeder, 2006, for reviews). For example, watching a silent video of an object that normally produces a sound (e.g., a clip of a dog's barking movements) induced activities in auditory cortex in humans; critically, different activation patterns were observed for different categories of visual stimuli (animals, musical instruments, and objects; Meyer et al., 2010; see also Fassnidge, Marcotti, & Freeman, 2017). In a similar vein, decoding the neural activity patterns in visual cortex when the participants were blindfolded can predict the category of the soundtrack (forest, traffic, or people) that they were listening to (Vetter, Smith, & Muckli, 2014). Such results therefore suggest that semantic information can be represented in the (presumed)

modality-specific areas even though the input was from another sensory modality. Such evidence does not, of course, mean that the convergence area of audiovisual semantic congruency is not necessary. Instead, it appears that audiovisual semantic information can be represented crossmodally in earlier stages of information processing than many people used to believe, suggesting an ambiguous boundary between perceptual and semantic systems (see also Ostarek & Huettig, 2017). Given the slow and rapid crossmodal semantic congruency effects that emerge as a function of the cue-target SOAs demonstrated in the present study, future research should try to understand how meaningful visual and auditory information interact dynamically in both modality-specific and multisensory brain areas.

Lexical-semantic interactions

According to Glaser and Glaser's (1989) model, the lexical representation activated by a linguistic stimulus contains only orthographic and phonological codes, while the associated meaning is stored in the semantic system. Accordingly, given that there is only a single semantic system, the semantic representation activated by a dog's barking sound and a spoken word "DOG", though not completely equivalent, should be highly similar.

Considering whether multisensory stimuli originate from a common source or not, this critical factor has led researchers to the suggestion of a different functional role for naturalistic sounds and spoken words: Naturalistic sounds and visual objects typically share a common source, so their associations are established and maintained by the statistical regularity of their co-occurrence. On the other hand, spoken words and visual objects originate from different sources, so their mappings are more arbitrary and likely involve higher-level cognitive/linguistic functions. Consequently, Waxman and Gelman (2009) proposed that naturalistic sounds are simply *associated* with objects, whereas spoken words *refer* to the concepts associated with the objects – a representation that is more abstract than a particular entity presented in a specific context. That is, as compared to a naturalistic sound that is taken as one of the features associating with a given object, a spoken word

links to a deeper conceptual representation that is more categorical and insensitive to the modality of the input signals (see also Lupyan & Thompson-Schill, 2012).

In the current study, the slow semantic congruency effect demonstrated that the presentation of a linguistic stimulus belonging to the basic-level category, either as a cue or as a target, led to the semantic interaction occurring at the basic level. In contrast, the congruency effect elicited by naturalistic sounds on visual pictures occurred at the superordinate level even though they are associated with basic-level objects. These results therefore suggest that linguistic and non-linguistic stimuli, when belonging to the basic level of the category hierarchy, may access semantic representations at different levels of the category hierarchy at the first step.

Taken together, it would seem that the lexical representation refers to a semantic representation precisely at the corresponding level of category hierarchy. This semantic representation is perhaps more conceptual or symbolic in nature than that associated with naturalistic sounds. This then leads to the conjecture of the necessity of a high-order convergence zone in semantic system linking to the lexical system – it can either be an amodal or a multisensory semantic system (the semantic representation of "DOG" in the oval with dashed outline in Figure 5). The non-linguistic stimuli (like visual pictures and naturalistic sounds), instead, are processed in a coarse-to-fine (or abstract-to-specific) fashion during semantic access. In addition, their lower-order semantic representations are likely modality specific, or simply grounded in the perceptual systems (the semantic representation of "dog" in the oval with solid line in Figure 5).

Conclusions

In the present study, we compared the time-courses of crossmodal semantic congruency effects by naturalistic sounds and spoken words on visual pictures and printed words. Three reliable effects were observed: a slow semantic congruency effect elicited by the presentation of a congruent comparing to incongruent cue, a rapid inhibitory effect elicited by the presentation of an incongruent cue comparing to noise, and a unique inhibitory effect elicited by the presentation of a

mismatched spoken word comparing to noise. By manipulating SOAs and semantic relatedness between cue and target, the results suggest that non-linguistic stimuli access semantic representations directly along coarse-to-fine levels of the category hierarchy; in contrast, linguistic stimuli access semantic representations at a particular level of the category hierarchy mediated by lexical representations. We have implemented such dynamic processing of each type of stimulus, as suggested in Glaser and Glaser's (1989) model, and therefore propose a more comprehensive framework in Figure 5. On the one hand, this new framework suggests that crossmodal semantic interactions can occur at multiple stages of human information processing, and plausibly extend all the way through to an early stage of visual and auditory processing. On the other hand, this framework exhibits the complexity of semantic processing in terms of the domains of category hierarchy (i.e., superordinate, basic, and subordinate levels) and convergence across modalities (i.e., from modality-specific to amodal/multisensory).

Acknowledgments

The authors were supported by the Arts and Humanities Research Council (AHRC), *Rethinking the Senses* grant (AH/L007053/1).

References

Adam, R., & Noppeney, U. (2010). Prior auditory information shapes visual category-selectivity in ventral occipito-temporal cortex. *NeuroImage, 52,* 1592-1602.

Arnal, L. H., Wyart, V., & Giraud, A. L. (2011). Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. *Nature Neuroscience, 14,* 797-801.

Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology, 63,* 1-29.

Ballas, J. A., & Howard, J. H. (1987). Interpreting the language of environmental sounds. *Environment and Behavior, 19,* 91-114.

Barsalou, L. W. (1999). Perceptions of perceptual symbols. *Behavioral and Brain Sciences, 22,* 637-660.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology, 59,* 617-645.

Barsalou, L. W. (2016). On staying grounded and avoiding quixotic dead ends. *Psychonomic Bulletin & Review, 23,* 1122-1142.

Bates, E., D'Amico, S., Jacobsen, T., Székely, A., Andonova, E., Devescovi, A., … Tzeng, O. (2003). Timed picture naming in seven languages. *Psychonomic Bulletin & Review, 10,* 344-380.

Bertelson, P., & Tisseyre, F. (1969). The time-course of preparation: Confirmatory results with visual and auditory warning signals. *Acta Psychologica, 30,* 145-154.

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences, 15,* 527-536.

Boutonnet, B., & Lupyan, G. (2015). Words jump-start vision: A label advantage in object recognition. *Journal of Neuroscience, 35,* 9329-9335.

Chen, Y.-C., & Spence, C. (2010). When hearing the bark helps to identify the dog: Semantically-congruent sounds modulate the identification of masked pictures. *Cognition, 114,* 389-404.

Chen, Y.-C., & Spence, C. (2011). Crossmodal semantic priming by naturalistic sounds and spoken words enhances visual sensitivity. *Journal of Experimental Psychology: Human Perception and Performance, 37,* 1554-1568.

Chen, Y.-C., & Spence, C. (2013). The time-course of the cross-modal semantic modulation of visual picture processing by naturalistic sounds and spoken words. *Multisensory Research, 26*, 371-386.

Chen, Y.-C., & Spence, C. (2017). Dissociating the time courses of the cross-modal semantic priming effects elicited by naturalistic sounds and spoken words. *Psychonomic Bulletin & Review*. doi: 10.3758/s13423-017-1324-6.

Clarke, A. (2015). Dynamic information processing states revealed through neurocognitive models of object semantics. *Language, Cognition and Neuroscience, 30,* 409-419.

Clarke, A., Devereux, B. J., Randall, B., & Tyler, L. K. (2015). Predicting the time course of individual objects with MEG. *Cerebral Cortex, 25,* 3602-3612.

Colavita, F. B. (1974). Human sensory dominance. *Perception & Psychophysics, 16,* 409-412.

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology, 1*, 42-45.

Cummings, A., Čeponienė, R., Koyama, A., Saygin, A. P., Townsend, J., & Dick, F. (2006). Auditory semantic networks for words and natural sounds. *Brain Research, 1115*, 92-107.

Damasio, A. R. (1989). Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition, 33*, 25-62.

de Haas, B., Schwarzkopf, D. S., Urner, M., & Rees, G. (2013). Auditory modulation of visual stimulus encoding in human retinotopic cortex. *NeuroImage, 70,* 258-267.

Dick, F., Saygin, A. P., Galati, G., Pitzalis, S., Bentrovato, S., D'Amico, S., ... & Pizzamiglio, L. (2007). What is involved and what is necessary for complex linguistic and nonlinguistic auditory processing: Evidence from functional magnetic resonance imaging and lesion data. *Journal of Cognitive Neuroscience, 19,* 799-816.

Doehrmann, O., & Naumer, M. J. (2008). Semantics and the multisensory brain: How meaning modulates processes of audio-visual integration. *Brain Research, 1242,* 136-150.

Donohue, S. E., Appelbaum, L. G., Park, C. J., Roberts, K. C., & Woldorff, M. G. (2013). Cross-modal stimulus conflict: The behavioral effects of stimulus input timing in a visual-auditory Stroop task. *PLoS ONE, 8*(4), e62802. doi:10.1371/journal.pone.0062802

Driver, J., & Noesselt, T. (2008). Multisensory interplay reveals crossmodal influences on 'sensory-specific' brain regions, neural responses, and judgments. *Neuron, 57,* 11-23.

Edmiston, P., & Lupyan, G. (2015). What makes words special? Words as unmotivated cues. *Cognition, 143*, 93-100.

Endress, A. D., & Potter, M. C. (2012). Early conceptual and linguistic processes operate in independent channels. *Psychological Science, 23,* 235-245.

Fabre-Thorpe, M. (2011). The characteristics and limits of rapid visual categorization. *Frontiers in Psychology, 2*:243. doi: 10.3389/fpsyg.2011.00243

Fassnidge, C., Marcotti, C. C., & Freeman, E. (2017). A deafening flash! Visual interference of auditory signal detection. *Consciousness and Cognition, 49,* 15-24.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175-191.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex, 1,* 1-47.

Fodor, J. A. (1975). *The language of thought*. Cambridge, MA: Harvard University Press.

Fox, L. A., Shor, R. E., & Steinman, R. J. (1971). Semantic gradients and interference in naming color, spatial direction, and numerosity. *Journal of Experimental Psychology, 91,* 59-65.

Gallese, V., & Lakoff, G. (2005). The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive Neuropsychology, 22,* 455-479.

Ghazanfar, A. A., & Schroeder, C. E. (2006). Is neocortex essentially multisensory? *Trends in Cognitive Sciences, 10,* 278-285.

Glaser, M. O., & Glaser, W. R. (1982). Time course analysis of the Stroop phenomenon. *Journal of Experimental Psychology: Human Perception and Performance, 8,* 875-894.

Glaser, W. R., & Düngelhoff, F. J. (1984). The time course of picture-word interference. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 640-654.

Glaser, W. R., & Glaser, M. O. (1989). Context effects in Stroop-like word and picture processing. *Journal of Experimental Psychology: General, 118,* 13-42.

Grill-Spector, K., & Kanwisher, N. (2005). Visual recognition: As soon as you know it is there, you know what it is. *Psychological Science, 16,* 152-160.

Hauk, O., Davis, M. H., Ford, M., Pulvermüller, F., & Marslen-Wilson, W. D. (2006). The time course of visual word recognition as revealed by linear regression analysis of ERP data. *NeuroImage, 30,* 1383-1400.

Hays, W. (1973). *Statistics for the social sciences.* New York, NY: Holt, Rinehart, & Winston.

Hocking, J., & Price, C. J. (2009). Dissociating verbal and nonverbal audiovisual object processing. *Brain and Language, 108,* 89-96.

Holcomb, P. J., & Grainger, J. (2006). On the time course of visual word recognition: An event-related potential investigation using masked repetition priming. *Journal of Cognitive Neuroscience, 18,* 1631-1643.

Iordanescu, L., Grabowecky, M., & Suzuki, S. (2011). Object-based auditory facilitation of visual search for pictures and words with frequent and rare targets. *Acta Psychologica, 137,* 252-259.

Iordanescu, L., Guzman-Martinez, E., Grabowecky, M., & Suzuki, S. (2008). Characteristic sounds facilitate visual search. *Psychonomic Bulletin & Review, 15,* 548-554.

Jolicoeur, P., Gluck, M. A., & Kosslyn, S. M. (1984). Pictures and names: Making the connection. *Cognitive Psychology, 16,* 243-275.

Kim, Y., Porter, A. M., & Goolkasian, P. (2014). Conceptual priming with pictures and environmental sounds. *Acta Psychologica, 146,* 73-83.

Klein, G. S. (1964). Semantic power measured through the interference of words with color-naming. *American Journal of Psychology, 77,* 576-588.

Koppen, C., Alsius, A., & Spence, C. (2008). Semantic congruency and the Colavita visual dominance effect. *Experimental Brain Research, 184,* 533-546.

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

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 207,* 203-205.

Lacouture, Y., & Cousineau, D. (2008). How to use MATLAB to fit the ex-Gaussian and other probability functions to a distribution of response times. *Tutorials in Quantitative Methods for Psychology, 4,* 35-45.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*:863. doi:10.3389/fpsyg.2013.00863.

Lakens, D. (2014). Calculating confidence intervals for Cohen's d and eta-squared using SPSS, R, and Stata. Retrieved from http://daniellakens.blogspot.co.uk/2014/06/calculating-confidence-intervals-for.html (29 Jun 2017).

Lambon Ralph, M. A., Jefferies, E., Patterson, K., & Rogers, T. T (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience, 18,* 42-55.

Lewis, J. W. (2010). Audio-visual perception of everyday natural objects - Hemodynamic studies in humans. In M. J. Naumer & J. Kaiser (Eds.), *Multisensory object perception in the primate brain* (pp. 155-190). New York, NY: Springer.

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology, 6:*1171. doi: 10.3389/fpsyg.2015.01171

Lupyan, G., & Thompson-Schill, S. L. (2012). The evocative power of words: Activation of concepts by verbal and nonverbal means. *Journal of Experimental Psychology: General, 141,* 170-186.

Mack, M.L., & Palmeri, T.J. (2015). The dynamics of categorization: Unraveling rapid categorization. *Journal of Experimental Psychology: General, 144,* 551-569.

Mädebach, A., Wöhner, S., Kieseler, M. L., & Jescheniak, J. D. (2017). Neighing, barking, and drumming horse – object related sounds help and hinder picture naming. *Journal of Experimental Psychology: Human Perception and Performance*. doi: 10.1037/xhp0000415.

Mahon, B. Z., & Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of Physiology-Paris, 102,* 59-70.

Mahon, B. Z., & Hickok, G. (2016). Arguments about the nature of concepts: Symbols, embodiment, and beyond. *Psychonomic Bulletin & Review, 23,* 941-958.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition, 25,* 71-102.

Meyer, A. S., & Schriefers, H. (1991). Phonological facilitation in picture-word interference experiments: Effects of stimulus onset asynchrony and types of interfering stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 1146-1160.

Meyer, K., Kaplan, J. T., Essex, R., Webber, C., Damasio, H., & Damasio, A. (2010). Predicting visual stimuli on the basis of activity in auditory cortices. *Nature Neuroscience, 13,* 667-668.

Marsolek, C. J. (2008). What antipriming reveals about priming. *Trends in Cognitive Sciences, 12,* 176-181.

Molholm, S., Ritter, W., Javitt, D. C., & Foxe, J. J. (2004). Multisensory visual-auditory object recognition in humans: A high-density electrical mapping study. *Cerebral Cortex, 14,* 452-465.

Murray, M. M., Camen, C., Andino, S. L. G., Bovet, P., & Clarke, S. (2006). Rapid brain discrimination of sounds of objects. *Journal of Neuroscience, 26,* 1293-1302.

Murray, M. M., Camen, C., Spierer, L., & Clarke, S. (2008). Plasticity in representations of environmental sounds revealed by electrical neuroimaging. *NeuroImage, 39,* 847-856.

Murray, M. M., & Spierer, L. (2009). Auditory spatio-temporal brain dynamics and their consequences for multisensory interactions in humans. *Hearing Research, 258,* 121-133.

Noppeney, U., Josephs, O., Hocking, J., Price, C. J., & Friston, K. J. (2008). The effect of prior visual information on recognition of speech and sounds. *Cerebral Cortex, 18,* 598-609.

Ostarek, M., & Huettig, F. (2017). Spoken words can make the invisible visible – Testing the involvement of low-level visual representations in spoken word processing. *Journal of Experimental Psychology: Human Perception and Performance 43,* 499-508.

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience, 8,* 976-987.

Plante, E., Van Petten, C., & Senkfor, A. J. (2000). Electrophysiological dissociation between verbal and nonverbal semantic processing in learning disabled adults. *Neuropsychologia, 38,* 1669-1684.

Plaut, D. C. (2002). Graded modality-specific specialisation in semantics: A computational account of optic aphasia. *Cognitive Neuropsychology, 19,* 603-639.

Pobric, G., Jefferies, E., & Lambon Ralph, M. A. (2010). Category-specific versus category-general semantic impairment induced by transcranial magnetic stimulation. *Current Biology, 20,* 964-968.

Potter, M. C. (2014). Detecting and remembering briefly presented pictures. In K. Kveraga & M. Bar (Eds.), *Scene vision: Making sense of what we see* (pp. 177-197). Cambridge, MA: MIT Press.

Pylyshyn, Z. W. (1984). *Computation and cognition*. Cambridge, MA: MIT Press.

Reilly, J., Peelle, J. E., Garcia, A., & Crutch, S. J. (2016). Linking somatic and symbolic representation in semantic memory: The dynamic multilevel reactivation framework. *Psychonomic Bulletin & Review, 23,* 1002-1014.

Robertson, I. H., Mattingley, J. B., Rorden, C., & Driver, J. (1998). Phasic alerting of neglect patients overcomes their spatial deficit in visual awareness. *Nature, 395,* 169-173.

Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition, 42,* 107-142.

Roelofs, A. (2005). The visual-auditory color-word Stroop asymmetry and its time course. *Memory & Cognition, 33,* 1325-1336.

Roelofs, A. (2010). Attention, temporal predictability, and the time course of context effects in naming performance. *Acta Psychologica, 133,* 146-153.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8,* 382-439.

Saygin, A. P., Dick, F., & Bates, E. (2005). An on-line task for contrasting auditory processing in the verbal and nonverbal domains and norms for younger and older adults. *Behavior Research Methods, 37,* 99-110.

Shimada, H. (1990). Effect of auditory presentation of words on color naming: The intermodal Stroop effect. *Perceptual and Motor Skills, 70,* 1155-1161.

Simanova, I., Hagoort, P., Oostenveld, R., & van Gerven, M. A. (2014). Modality-independent decoding of semantic information from the human brain. *Cerebral Cortex, 24,* 426-434.

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement, 61,* 605-632.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 174-215.

Snyder, J. S., & Gregg, M. K. (2011). Memory for sound, with an ear toward hearing in complex auditory scenes. *Attention, Perception, & Psychophysics, 73,* 1993-2007.

Soemer, A., & Saito, S. (2015). Maintenance of auditory-nonverbal information in working memory. *Psychonomic Bulletin & Review, 22,* 1777-1783.

Spence, C., Parise, C., & Chen, Y.-C. (2011). The Colavita visual dominance effect. In M. M. Murray & M. Wallace (Eds.), *Frontiers in the neural bases of multisensory processes* (pp. 529-556). Boca Raton, FL: CRC Press.

Stuart, D. M., & Carrasco, M. (1993). Semantic component of a cross-modal Stroop-like task. *American Journal of Psychology, 106,* 383-405.

Suied, C., Bonneel, N., & Viaud-Delmon, I. (2009). Integration of auditory and visual information in the recognition of realistic objects. *Experimental Brain Research, 194,* 91-102.

Taylor, K. I., Moss, H. E., Stamatakis, E. A., & Tyler, L. K. (2006). Binding crossmodal object features in perirhinal cortex. *Proceedings of the National Academy of Sciences of the U.S.A., 103,* 8239-8244.

Thierry, G., Giraud, A. L., & Price, C. (2003). Hemispheric dissociation in access to the human semantic system. *Neuron, 38,* 499-506.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381,* 520-522.

van Petten, C., Coulson, S., Rubin, S., Plante, E., & Parks, M. (1999). Time course of word identification and semantic integration in spoken language. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 394-417.

van Petten, C., & Rheinfelder, H. (1995). Conceptual relationships between spoken words and environmental sounds: Event-related brain potential measures. *Neuropsychologia, 33,* 485-508.

VanRullen, R., & Thorpe, S. J. (2001). The time course of visual processing: From early perception to decision-making. *Journal of Cognitive Neuroscience, 13,* 454-461.

Vetter, P., Smith, F. W., & Muckli, L. (2014). Decoding sound and imagery content in early visual cortex. *Current Biology, 24,* 1256-1262.

Waxman, S. R., & Gelman, S. A. (2009). Early word-learning entails reference, not merely associations. *Trends in Cognitive Sciences, 13,* 258-263.

Weatherford, K., Mills, M., Porter, A. M., & Goolkasian, P. (2015). Target categorization with primes that vary in both congruency and sense modality. *Frontiers in Psychology, 6*:20. doi:10.3389/fpsyg.2015.00020

Werner, S., & Noppeney, U. (2010). Superadditive responses in superior temporal sulcus predict audiovisual benefits in object categorization. *Cerebral Cortex, 20,* 1829-1842.

Wu, C. T., Crouzet, S. M., Thorpe, S. J., & Fabre-Thorpe, M. (2014). At 120 msec you can spot the animal but you don't yet know it's a dog. *Journal of Cognitive Neuroscience, 27,* 141-149.

Wuensch, K. L. (2016). Placing a confidence interval on multiple R2. Retrieved from http://core.ecu.edu/psyc/wuenschk/StatHelp/CI-R2.htm (29 Jun 2017).

Yuval-Greenberg, S., & Deouell, L. Y. (2009). The dog's meow: Asymmetrical interaction in cross-modal object recognition. *Experimental Brain Research, 193,* 603-614.

Figure Legends

Figure 1. The model used to account for the Stroop-like effects between visual pictures and printed words that was proposed by Glaser and Glaser (1989). There are two main cognitive systems in the model: Semantic memory stores knowledge concerning the objects in the outside world. The lexicon stores linguistic knowledge concerning the orthography and phonology of words, though it has no semantic capacity. These two cognitive systems have bidirectional connections, such that the meaning of a word needs to be retrieved from semantic memory, while the name of an object needs to be retrieved from the lexicon. The executive systems control perception and action regarding a particular type of stimulus or response. Specifically, each executive system receives sensory inputs and processes them, retrieves information from semantic memory and from the lexicon, and generates an appropriate response in order to fulfill the task at hand. Figure reproduced from Glaser and Glaser (1989, Figure 5a).

Figure 2. The sequence of stimuli presented in each trial: A blank screen was followed by a target picture (e.g., dog) presented until a response was made (or 2000 ms had elapsed). The timeline represents the seven possible SOAs between the onset of the auditory cue and the visual target. The auditory cue was either a naturalistic sound or a spoken word, and the relationship between the cue and target was congruent (i.e., the two stimuli are associated with the same object), related (i.e., they were both in the task-relevant category of living or non-living things), incongruent (i.e., one of the stimuli was a living thing and the other, a non-living thing), or else a white noise was presented as the neutral condition.

Figure 3. The median RTs in the congruent, related, incongruent, noise, and no-sound conditions in Experiment 1 where the visual targets were pictures, and the auditory cues were (A) naturalistic sounds or (B) spoken words. Error bars indicate ±1 standard error of the means in the within-subject design (Cousineau, 2005). Under the SOA axis, the red solid

line represents the time-window of semantic congruency effect, the blue dotted line represents the time-window of inhibition by the incongruent cue, and the green dashed line, the time-window of inhibition by the related cue. At each SOA, the significant pairwise t-tests are summarized with the label of each condition (C: congruent, R: related, I: incongruent, N: noise; see Table 3 for details).

Figure 4. The median RTs in the congruent, related, incongruent, noise, and no-sound conditions in Experiment 2 where the visual targets were printed words, and the auditory cues were (A) naturalistic sounds or (B) spoken words. Error bars indicate ±1 standard error of the means in the within-subject design (Cousineau, 2005). Under the SOA axis, the red solid line represents the time-window of semantic congruency effect, the blue dotted line represents the time-window of inhibition by the incongruent cue, and the green dashed line, the time-window of inhibition by the related cue. At each SOA, the significant pairwise t-tests are summarized with the label of each condition (C: congruent, R: related, I: incongruent, N: noise; see Table 5 for details).

Figure 5. Proposed framework for the crossmodal semantic interactions between linguistic and non-linguistic stimuli presented in the visual and auditory modalities. The framework structure adopts the key feature in Glaser and Glaser's (1989) model that the perceptual/semantic and lexical systems are separate but extensively connected (the separation and the long bidirectional connection in the figure simply aiming to clearly demonstrate the structure of each system). Based on the results of the present study and recent research, the following modifications are included in the current model: First, the top level represents the separate sensory processing in the visual and auditory modality. Second, given the modern view of the ambiguous boundary between perception and cognition, and recent evidence demonstrating that semantic information is available at the perceptual stage of information processing, the semantic system is grounded on perceptual processing to a certain extent. Third, the category hierarchy was plotted in the

semantic system in order to explain our results showing that the semantic interactions between visual pictures and naturalistic sounds occurred at the superordinate level. On the other hand, when either one, or both, of the cue and target were linguistic stimuli, their interaction occurred at the level of category hierarchy to which the word corresponds. In this example, the word "dog" corresponds to a basic-level category, and therefore the interactions associated with the word "dog" occur at the basic level. Finally, in the semantic system, it is possible that non-linguistic stimuli (visual pictures and naturalistic sounds) access semantic representations that are modality-specific (represented by the oval with a solid outline), whereas linguistic stimuli (printed and spoken words) access semantic representations that can be either amodal or multisensory (represented by the oval with a dashed outline).

Table 1. Possible hypotheses and mechanisms underlying audiovisual semantic interactions and the expected patterns of reaction times (RTs) in the congruent (cue and target are associated with the same object), related (both cue and target are living or non-living things), and incongruent (either cue or target belong to living thing, while the other, non-living thing) conditions.

Hypothesis	Mechanism	The pattern of RTs in the three conditions		
		Congruent	Related	Incongruent
Semantic congruency at the basic level	Stronger facilitation is elicited by cue on the processing of target when they have higher semantic relatedness	<	<	
Semantic congruency at the superordinate level	Cue facilitates the processing of target when they both belong to living- (or non-living-) thing category	=	<	
Lexical matching	Cue facilitates the processing of target through lexical-semantic interaction only when they share the same name	<	=	
Response compatibility	Cue speeds up the reaction time when cue and target map to the same response	=	<	

Table 2. Mean accuracy (%) in the picture categorization task in each condition (Experiment 1). Negative SOAs indicate that the auditory cue was presented first, while positive SOAs indicate that the target picture was presented first.

		SOA (ms)							
Sound type	Congruency	-1000	-500	-250	-100	0	100	250	Mean
Naturalistic sounds	Congruent	97.9	98.0	98.3	97.9	98.2	98.1	98.0	98.1
	Related	98.3	98.1	98.0	96.8	98.9	96.9	97.5	97.8
	Incongruent	96.8	97.6	98.1	97.0	98.0	98.3	98.4	97.7
	Noise	97.9	98.3	97.3	97.2	96.8	97.7	98.1	97.6
Spoken words	Congruent	97.9	98.5	96.9	97.6	97.6	97.7	98.5	97.8
	Related	96.8	97.0	98.4	98.3	98.3	97.7	98.5	97.9
	Incongruent	97.7	97.1	96.7	97.9	97.7	97.5	97.3	97.4
	Noise	97.7	97.5	97.1	97.2	97.7	97.9	98.1	97.6

Table 3. Experiment 1: The simple main effect analysis of median RTs on the Congruency factor at each SOA and the post-hoc pairwise t-tests. The t-values are reported, and the significance was based on the conditions of $df = 29$, two-tailed, and Holm-Bonferroni correction (*: $p < .05$; **: $p < .01$). Negative SOAs indicate that the auditory cue was presented first, while positive SOAs indicate that the target picture was presented first.

	Sound type	(A) Naturalistic sounds							(B) Spoken words						
	SOA (ms)	-1000	-500	-250	-100	0	100	250	-1000	-500	-250	-100	0	100	250
Simple main effect	*F*(3,87)	1.15	6.00	5.04	1.96	0.96	0.33	1.17	15.48	28.80	13.74	3.78	4.28	0.89	0.90
	p	0.33	< .005	< .005	0.13	0.42	0.80	0.33	< .001	< .001	< .001	< .05	< .01	0.45	0.45
Post-hoc t-tests	Congruent vs. Incongruent		3.27 *	3.60 **					5.64 **	9.10 **	5.90 **				
	Congruent vs. Noise								6.39 **	6.37 **					
	Related vs. Noise								3.39 **			3.36 *	2.94 *		
	Incongruent vs. Noise			2.92 *							4.18 **	2.77 *	3.48 *		
	Congruent vs. Related								3.77 **	4.68 **					
	Related vs. Incongruent		3.59 **							4.80 **	3.48 **				

Table 4. Mean accuracy (%) in the word categorization task in each condition (Experiment 2). Negative SOAs indicate that the auditory cue was presented first, while positive SOAs indicate that the target word was presented first.

		SOA (ms)							
Sound type	Congruency	-1000	-500	-250	-100	0	100	250	Mean
Naturalistic sounds	Congruent	98.0	97.6	98.4	98.0	97.7	98.1	97.6	97.9
	Related	95.9	97.7	97.3	97.5	97.9	98.4	97.1	97.4
	Incongruent	97.7	94.9	95.7	94.8	98.5	97.4	98.3	96.8
	Noise	97.5	97.3	97.7	97.1	96.1	96.9	97.5	97.2
Spoken words	Congruent	95.8	97.1	97.2	97.0	98.3	97.1	96.9	97.5
	Related	96.8	97.3	97.7	98.4	96.7	97.9	97.3	97.4
	Incongruent	97.1	95.7	96.5	96.8	97.0	96.9	98.0	96.8
	Noise	97.6	98.1	97.3	95.9	96.7	96.4	96.0	97.0

Table 5. Experiment 2: The simple main effect analysis of median RTs on the Congruency factor at each SOA and the post-hoc pairwise t-tests. The t-values are reported, and the significance was based on the conditions of $df = 29$, two-tailed, and Holm-Bonferroni correction (*: $p < .05$; **: $p < .01$). Negative SOAs indicate that the auditory cue was presented first, while positive SOAs indicate that the target word was presented first.

	Sound type	(A) Naturalistic sounds							(B) Spoken words						
	SOA (ms)	-1000	-500	-250	-100	0	100	250	-1000	-500	-250	-100	0	100	250
Simple main effect	*F*(3,87)	1.34	12.69	14.24	3.00	3.91	1.49	1.73	12.45	14.33	8.42	6.38	10.16	5.13	3.16
	p	0.27	< .001	< .001	< .05	< .05	0.22	0.17	< .001	< .001	< .001	< .005	< .001	< .005	< .05
Post-hoc t-tests	Congruent vs. Incongruent		5.14 **	6.03 **					4.81 **	5.33 **	4.69 **	3.92 **	3.20 *		
	Congruent vs. Noise		4.24 **	3.54 **					4.60 **	6.57 **	2.59 *				
	Related vs. Noise												4.09 **		
	Incongruent vs. Noise			3.68 **	2.95 *	3.26 *						2.99 *	4.74 **	4.26 **	
	Congruent vs. Related		4.00 **	4.28 **					4.07 **	5.55 **	2.66 *	2.75 *			
	Related vs. Incongruent										3.76 **				

Table 6. Summary of the crossmodal semantic congruency effects reported in Experiments 1 and 2.

Auditory cue	Visual target	Slow congruency effect (congruent vs. incongruent condition)	Rapid inhibitory effect (incongruent vs. noise condition)	Inhibitory effect by spoken words (related vs. noise or congruent condition)
Naturalistic sounds	Picture	-500 ms SOA Superordinate level	-250 ms SOA	N/A
	Word	-500 to -250 ms SOA Basic level	-250 to 0 ms SOA	N/A
Spoken words	Picture	-1000 to -500 ms SOA Basic level	-250 to 0 ms SOA	-100 to 0 ms SOA
	Word	-1000 to -250 ms SOA Basic level	-100 to 100 ms SOA	-100 to 0 ms SOA

Figure

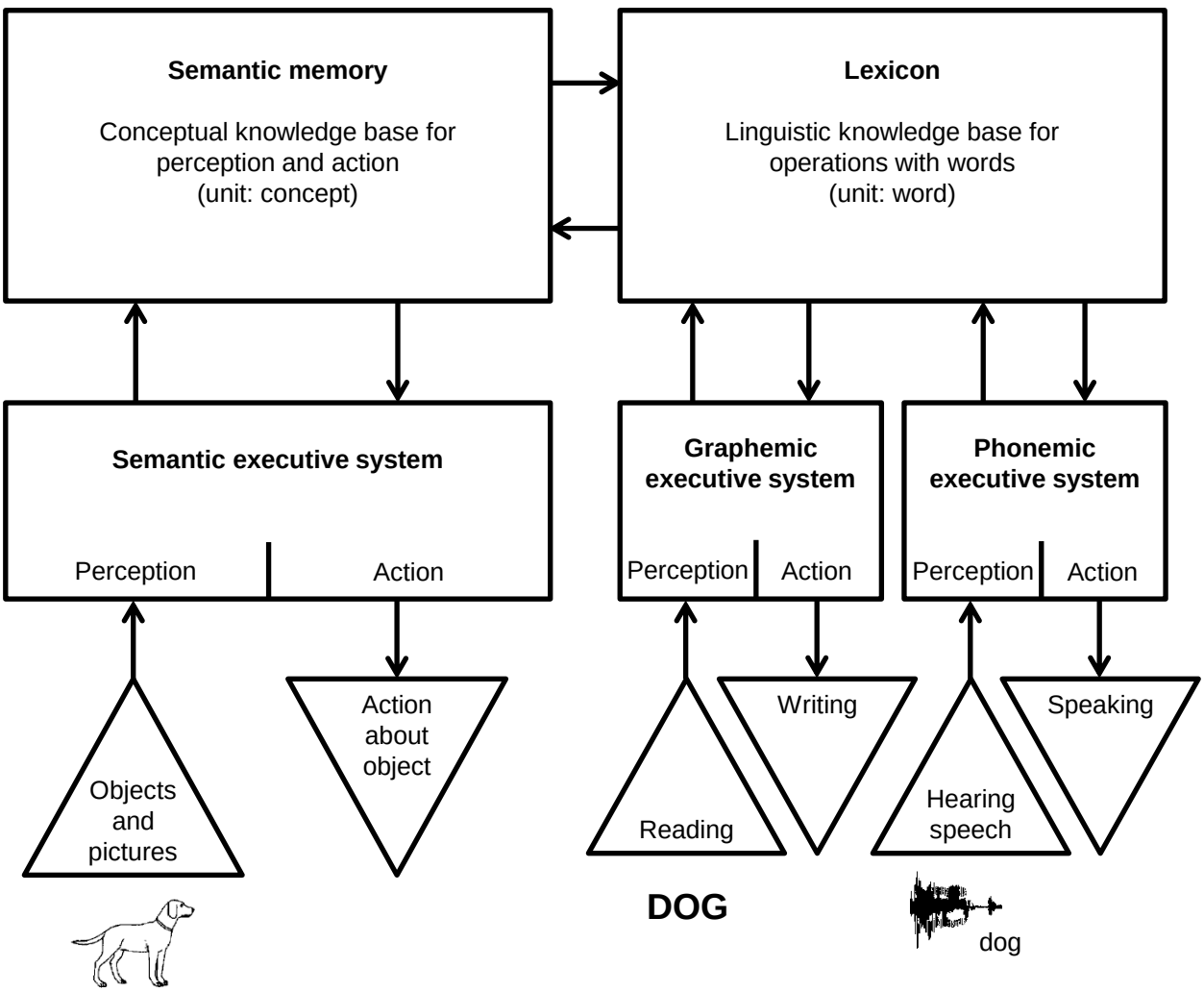
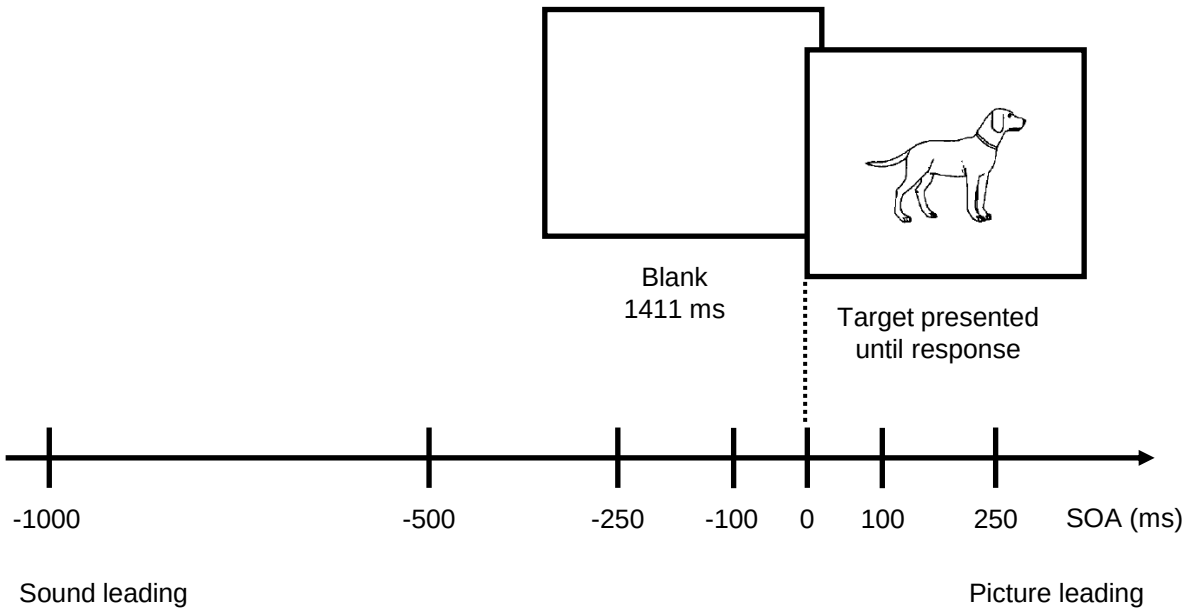


Figure 1


Blank
1411 ms
Target presented
until response
-1000
-500
-250
-100
0
100
250
SOA (ms)
Sound leading
Picture leading

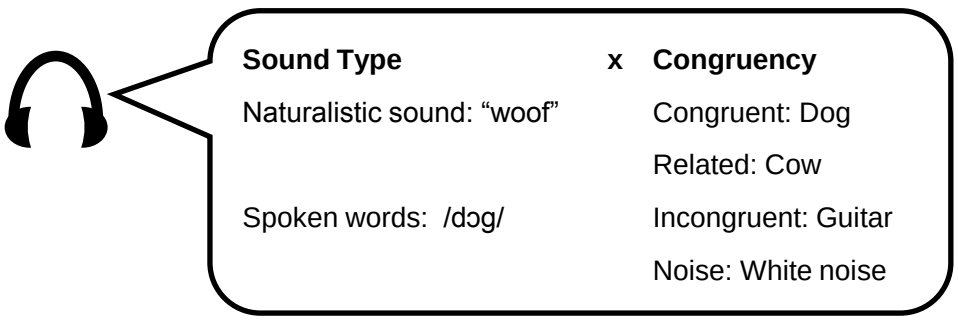

Sound Type
x
Congruency
Naturalistic sound: “woof”
Congruent: Dog
Related: Cow
Spoken words: /dɔg/
Incongruent: Guitar
Noise: White noise

Figure 2

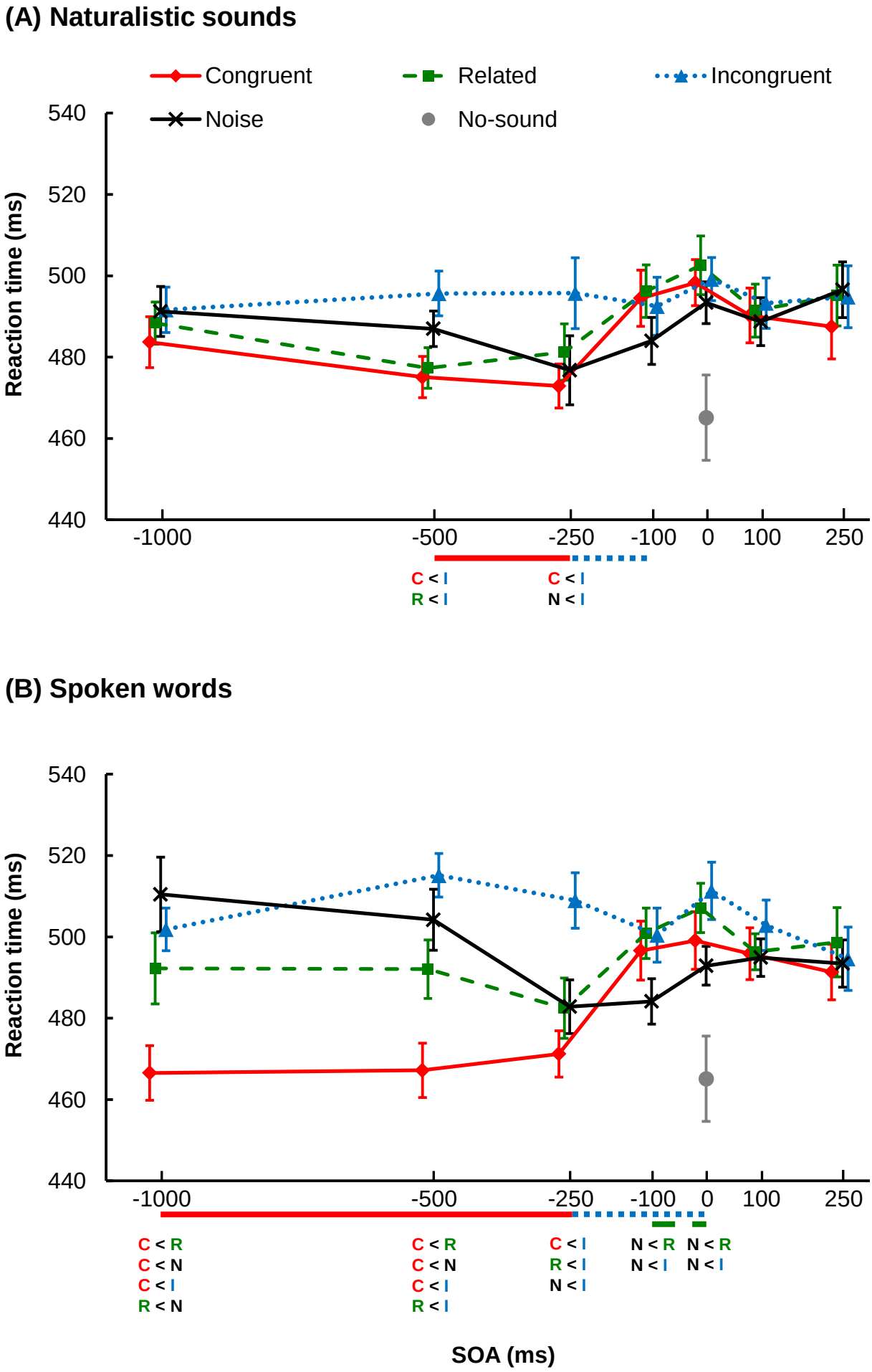

(A) Naturalistic sounds
Congruent
Related
Incongruent
Noise
No-sound
Reaction time (ms)
540
520
500
480
460
440
-1000
-500
-250
-100
0
100
250
C < I
R < I
C < I
N < I
(B) Spoken words
Reaction time (ms)
540
520
500
480
460
440
-1000
-500
-250
-100
0
100
250
C < R
C < N
C < I
R < N
C < R
C < N
C < I
R < I
C < I
R < I
N < I
N < R
N < I
N < R
N < I
SOA (ms)

Figure 3

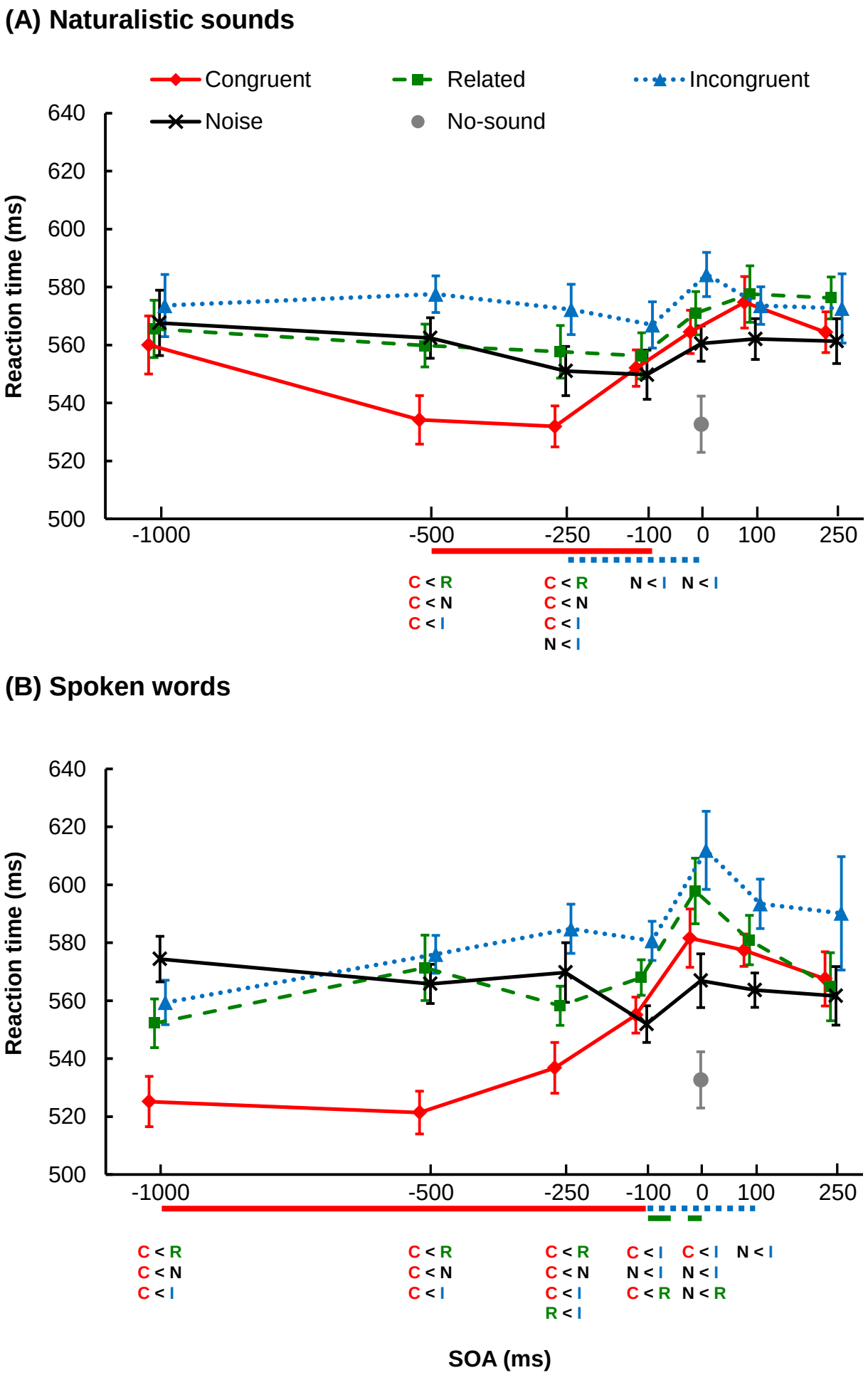

(A) Naturalistic sounds
Congruent
Related
Incongruent
Noise
No-sound
Reaction time (ms)
640
620
600
580
560
540
520
500
-1000
-500
-250
-100
0
100
250
C < R
C < N
C < I
C < R
C < N
C < I
N < I
N < I
N < I
(B) Spoken words
Reaction time (ms)
640
620
600
580
560
540
520
500
-1000
-500
-250
-100
0
100
250
C < R
C < N
C < I
C < R
C < N
C < I
C < R
C < N
C < I
R < I
C < I
N < I
C < R
C < I
N < I
N < R
N < I
SOA (ms)

Figure 4

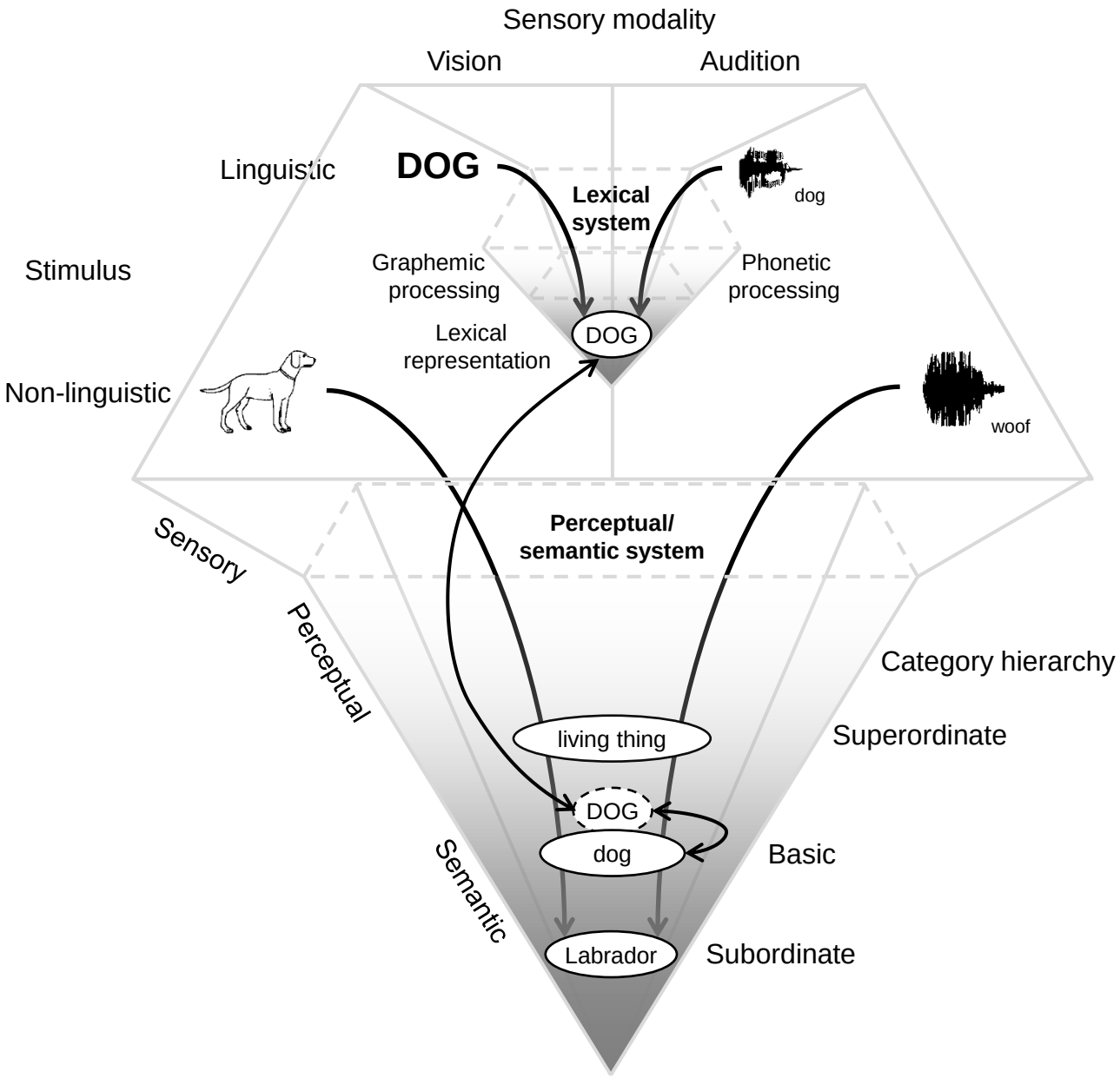

Sensory modality
Vision
Audition
Linguistic
DOG
dog
Lexical system
Stimulus
Graphemic processing
Phonetic processing
DOG
Lexical representation
Non-linguistic
woof
Perceptual/ semantic system
Sensory
Perceptual
Category hierarchy
living thing
Superordinate
DOG
dog
Basic
Semantic
Labrador
Subordinate

Figure 5