LANGUAGE, COGNITION AND NEUROSCIENCE
2020, VOL. 35, NO. 10, 1456–1464
https://doi.org/10.1080/23273798.2020.1797838

Routledge
Taylor & Francis Group

REGULAR ARTICLE

Degree of conceptual overlap affects eye movements in visual world paradigm

Haibin Han[a,b] and Xingshan Li[a]

[a]CAS Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences, Beijing, People's Republic of China;
[b]Department of Psychology, University of Chinese Academy of Sciences, Beijing, People's Republic of China

ABSTRACT

Several studies employing the visual world paradigm have shown that people will look at visual objects having a semantic relationship with spoken words. However, it remains unclear how eye movements are controlled in the visual world paradigm. The present study examined how listeners guide their eyes in the visual world paradigm by distinguishing the conceptual overlap hypothesis and the uncertainty reduction hypothesis. The uncertainty of the spoken words and the degree of conceptual overlap between visual objects and spoken words were manipulated by varying the spoken words across different hierarchical levels. The result showed that participants looked at the target objects more often when there was greater conceptual overlap between visual objects and spoken words, suggesting that the degree of conceptual overlap between spoken words and visual objects is the major factor that determines the probability of looking at a target object.

ARTICLE HISTORY
Received 31 August 2019
Accepted 11 July 2020

KEYWORDS
Visual attention; conceptual overlap; visual world paradigm; eye movements; spoken language comprehension

Using the visual world paradigm pioneered by Cooper (1974) and Tanenhaus et al. (1995), previous studies have shown that our eyes usually look at objects mentioned in the speech we hear (Altmann & Kamide, 2007; Huettig & Altmann, 2005; Yee & Sedivy, 2001). The visual world paradigm clearly presents how to investigate the integration of language and visual information, which is important in day-to-day activities, where people relate what they hear to what they see. In this paradigm, auditory stimuli are presented to participants along with a visual display in which a scene with different objects is depicted. The results show that visual attention is directed not only to objects directly mentioned in speech but also to those that are semantically related to the spoken words. These results raise an interesting question of what drives eye movements when people integrate speech and visual information.

Some studies have suggested that the probability of viewing a given object is proportional to the extent of the conceptual overlap between the object and the spoken word (Allopenna et al., 1998; Altmann & Kamide, 2007; Huettig & Altmann, 2005; Tanenhaus et al., 2000). For example, Huettig and Altmann (2005) found that participants directly looked at a *trumpet* when they heard a semantically related spoken word, in this case, *piano*. They concluded that "hearing 'piano' activated semantic information which overlapped with the semantic information encoded within the mental representation of the concurrent 'trumpet'" (p. B30). This suggests that eye movements are driven by the degree of conceptual overlap (both items were musical instruments). In another study, rather than using the semantic relation between the visual scene and spoken words, Myung et al. (2006) considered a specific feature (how to manipulate an object) as the overlapping feature and showed that common functional features shared by visual objects and spoken words (e.g. *piano* and *typewriter*), could also increase looking toward the typewriter than toward unrelated objects, such as a *bucket*. In their manipulation, they considered the common feature of a *piano* and a *typewriter*, which is that their use requires similar hand positions and movements.

Based on the findings reviewed above, Altmann and Kamide (2007) formally proposed a *conceptual overlap* account to explain why linguistic information influences the probability that people shift their visual attention in studies using the visual world paradigm. This theory makes two assumptions. First, the conceptual representations of spoken words and objects are compositions of features rather than indivisible wholes, and each object usually has several criteria that distinguish it from other objects. Second, the activation of conceptual features by spoken words increases the preexisting activation of conceptual representations of visual objects, which is called an *activation boost* caused by featural overlap. Based on these assumptions, Altmann and Kamide (2007) proposed that the greater the featural

CONTACT Xingshan Li lixs@psych.ac.cn

overlap between the object shown in a picture and the object mentioned in speech, the greater the likelihood of a saccade toward the target. However, although many studies have shown that conceptual overlap can mediate visual attention, no study has provided strong evidence by directly manipulating the degree of conceptual overlap between spoken words and visual objects.

An alternative hypothesis to the conceptual overlap hypothesis is the *uncertainty reduction hypothesis*. According to this hypothesis, one is more likely to look at a visual object which helps to reduce the degree of uncertainty of the spoken language. That is to say, when the spoken word is more uncertain, listeners might have difficulty comprehending word. In this situation, listeners might need to look at the corresponding visual object so that they can understand the spoken language more efficiently. In contrast, if the spoken language is specific and informative, listeners can understand the spoken word only with spoken language, so that they do not necessarily look at the corresponding visual object. For instance, when we hear the word *animal*, we do not know exactly what kind of animal the speaker is referring to, and thus listeners need to look at the corresponding visual object to reduce the uncertainty of the word *animal*. On the contrary, a more specific spoken word such as a *sparrow* is less uncertain, so that listeners can comprehend the word without looking at the corresponding visual object; thus, listeners may not necessarily look at the corresponding visual object. In summary, the uncertainty reduction hypothesis predicts that listeners are more likely to look at the corresponding visual object if a spoken word is more uncertain.

The current study was designed to understand how listeners guide their eyes in the visual world paradigm by distinguishing the conceptual overlap hypothesis and the uncertainty reduction hypothesis. We asked participants to listen to three different categories of spoken words while looking at a visual scene with four objects. The spoken words were manipulated in such a way that they represented different concepts of different hierarchical levels. For example, when people see a picture as shown in Figure 1, which includes a *sparrow*, they may hear a spoken keyword of either *animal*, *bird,* or *sparrow*. Traditionally, the concepts can be divided into three levels based on the abstraction level of the concept: general and abstract superordinate level, intermediate basic level, and more specific subordinate level (Rosch et al., 1976). These three hierarchical levels of concepts differ in the number of features shared by their members. The features of superordinate-level members are highly distinct and lack specificity, subordinate-level members have very specific features, and the basic-level members balanced the other two. As we will explain afterward, the words corresponding to different levels of concepts differ based on the degree of conceptual overlap between spoken words and visual objects, and also differ in terms of the uncertainty. Because

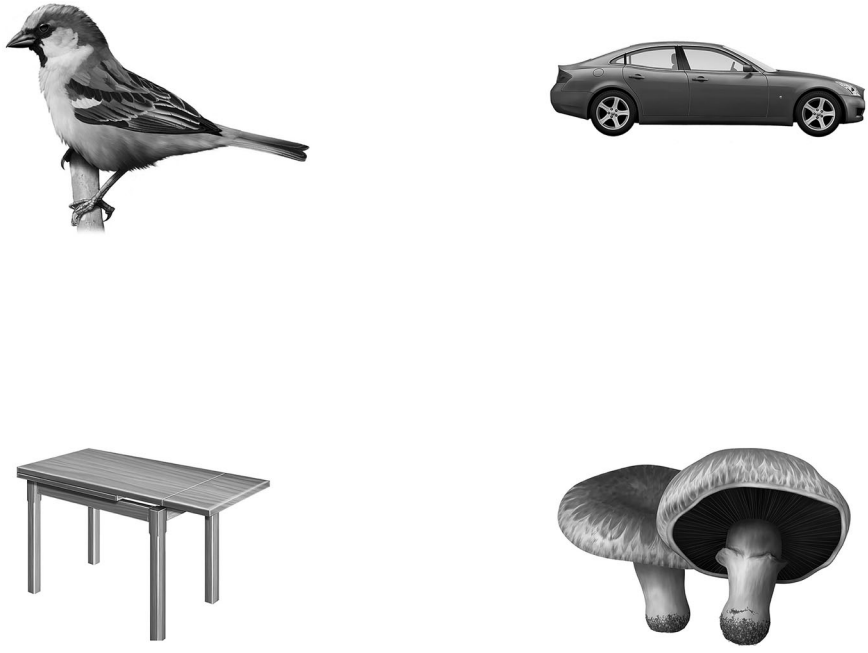

Figure 1. An example of a visual display used in the present study.

Note: In this figure, the target is a sparrow and the picture of the car, the table, and the mushroom are distractors. Across all the visual sets, the positioning of the target and the distractors was random.

different papers have different descriptions of the conceptual level, and we were interested in the comparison of uncertainty and degree of overlaps between spoken words and visual objects, we refer to the words corresponding to the highest, lowest and middle level concepts as low-specific words (e.g. animals), high-specific words (e.g. sparrow), and medium-specific words (e.g. bird), respectively.

On one hand, such manipulation can meet the needs of using spoken words at different hierarchical levels to investigate how the degree of conceptual overlap affects eye movements. A visual scene with four objects was presented to the participants. A high-specific word (e.g. *Labrador*, *sparrow*) has the most specific and informative features, and these features were shared by the target object in the visual display (e.g. a picture of a *sparrow*), thereby a high-specific word has the highest degree of conceptual overlap with a visual target object. In contrast, a low-specific word (e.g. *animal*, *furniture*) carry the fewest conceptual features, and thus they have the lowest degree of conceptual overlap with visual objects. Similarly, the degree of conceptual overlap between a medium-specific word (e.g. *bird*, *dog*) and the target object lies between the above two conditions. According to the conceptual overlap hypothesis proposed by Altmann and Kamide (2007), fewer overlapping features between the objects shown in the visual scene and the words in speech result in fewer looks at the targets. The hypothesis predicts that the target objects will be looked at most, an intermediately, and least when participants hear the high-specific, medium-specific, and low-specific words, respectively.

On the other hand, as the spoken words presented to the participants were selected from different hierarchical levels based on their levels of abstraction, we can also examine whether the uncertainty of the spoken words affects our eye movements. In particular, low-specific words at a high hierarchical level were more abstract and had more referents (e.g. *animal*). In contrast, high-specific words at the lowest hierarchical level were less abstract and had few referents (e.g. *sparrow*) and medium-specific words at an intermediate hierarchical level had a balanced position (e.g. *bird*). Consequently, the uncertainty of the low-specific spoken words is higher than that of medium-specific words, which is in turn higher than that of high-specific words. Based on the proposition of the uncertainty reduction hypothesis, when a listener hears a word with higher uncertainty, the eyes are more likely to move to the relevant object to assist in spoken language comprehension. Thus, contrary to the prediction of the conceptual overlap hypothesis, the uncertainty reduction hypothesis predicts that readers' eyes are more likely to move to the relevant visual object when they hear a low-specific word (e.g. animal) than when they hear a medium-specific word (e.g. bird), and when they hear a medium-specific word (e.g. bird) than when they hear a high-specific word (e.g. sparrow).

The present study could be used to distinguish between the conceptual overlap hypothesis and uncertainty reduction hypothesis regarding how eye movements are guided in visual world paradigm experiments. We were mainly interested in the following two eye movement behaviours: First, the proportion of eye movements that are ultimately attracted to the objects mentioned by different-level words; second, the speed on which the spoken words from the different levels mapped to the objects. These eye movement behaviours can provide us a good way to observe the dynamic aspect of visual attention deployment during visual-language integration.

Method

Participants

A total of 42 native Chinese speakers, aged between 18 and 29 years ($M = 23.2$) participated in the experiment.

Apparatus

The eye movements of the participants were recorded using an eye tracker with a sample rate of 1000 Hz. Experimental materials were presented on a 21-inch CRT monitor (SONY G520) with a 1024×768 pixel resolution and a refresh rate of 150 Hz. Although the viewing was binocular, only the right-eye movement data were collected. The participants were seated at a distance of 58 cm from the video monitor.

Materials and design

A total of 42 sets of critical visual displays were each paired with a spoken word for the experiment. Each visual display contained a target (e.g. *sparrow*) and three distractors (see Figure 1 for an example of the experimental stimuli). Three conditions were created depending on the hierarchical level of the spoken words. In the high-specific condition, the spoken words were at the lowest hierarchical level, sharing more common features with the visual object (e.g. "麻雀", /ma2que4/, meaning *sparrow*); in the medium-specific condition, the spoken words were at the medium hierarchical level (e.g. "鸟", /niao3/, *bird*), and in the low-specific condition, the spoken words were at the

highest hierarchical level (e.g. "动物", /dong4wu4/, *animal*). We selected 42, 8, and 6 words as high-specific, medium-specific, and low-specific words, respectively. These words were chosen from Battig and Montague (1969) and Rosch et al. (1976) and were translated into Chinese for the experiment.

The degree of matchiness between the pictures and the spoken words was evaluated using a picture-name agreement task. We carefully ensured that the distractors and the spoken target words do not share any phonological, semantic, or orthographic components. The positions of the four objects in the visual scenes were fully counterbalanced in the display.

Furthermore, medium- and low-specific words appeared two or more times in the experiment since they were fewer than high-specific words. To reduce the influence of these repeated items, 42 additional fillers from the medium-specific and high-specific conditions were added so that not only medium-specific and low-specific words but also high-specific ones were repeated. Each filler trial contained four objects unrelated to any spoken word. At the beginning of the experiment, eight practice trials were conducted to familiarise the participants with the procedure. The spoken words were recorded by a female native speaker of Chinese, and the mean duration of the spoken words was comparable ($ps > .153$) across three conditions (high-specific condition: $M = 765.23$, $SD = 59.43$; medium-specific condition: $M = 735.56$, $SD = 122.92$; Low-specific condition: $M = 716.15$, $SD = 31.15$).

The pictures used in the present study were selected from the Merriam-Webster visual dictionary, designed by QA International. All pictures were black and white and adjusted to a size of 350×250 pixels, subtending an approximate $13.67° \times 9.77°$ visual angle, to fit the experimental computer monitor. To measure the visual complexity of the pictures, 15 participants were instructed to rate the complexity of the picture of each object on a 7-point scale, for which 1 indicated very simple and 7 indicated very complex. The complexity of the pictures was carefully controlled. The participants were told to rate the complexity of the drawings themselves rather than the complexity of the real-life objects they represented. Visual complexity was comparable across the targets ($M = 3.987$, $SD = 1.331$) and distractor objects (distractor 1: $M = 34.179$, $SD = 1.082$; distractor 2: $M = 33.841$, $SD = 1.024$; distractor 3: $M = 33.685$, $SD = 1.156$), with no significant differences displayed ($ps > .248$). To measure luminance, we used the function "*lum_calc*", which was designed by Rodrigo Dal Ben (2019). The function uses the *MATLAB Image Processing Toolbox* to transform RGB images into HSV and CIE Lab colour spaces. From the value and luminance channels, the luminance mean and standard deviation were calculated. The mean luminance was comparable across target and distractor objects. The luminance difference from the HSV colour space between targets ($M = 0.874$, $SD = 0.228$) and distractors (distractor 1: $M = 0.860$, $SD = 0.236$ distractor 2: $M = 0.852$, $SD = 0.227$, $p = 0.269$; distractor 3: $M = 0.854$, $SD = 0.235$, $p = .321$) were not significant ($ps > .269$). The luminance difference from the CIE Lab colour space between targets ($M = 87.945$, $SD = 21.980$) and distractors (distractor 1: $M = 86.622$, $SD = 22.823$; distractor 2: $M = 85.926$, $SD = 21.791$; distractor 3: $M = 86.042$, $SD = 22.709$) was also not significant ($ps > .286$).

Procedure

The eye tracker was calibrated and validated at the beginning of the experiment. The validation error was smaller than 1° of the visual angle. The visual display was presented for two seconds before the onset of the spoken word. The participants were instructed to press the left button if any objects on the visual display matched the spoken word, and to press the right button otherwise.

Results

Reaction times (*RTs*) and the accuracies of the button press were calculated for the three conditions. Using the eye movement data, we calculated the mean fixation proportions on the targets and distractors under each condition from 500 ms before to 1,500 ms after the onset of the spoken target word. The period was divided into 21-time windows, with the fixation proportions of each object on the display being measured every 100 ms.

Both the behaviour and eye movement data were analysed using logit mixed models (Jaeger, 2008), in which word type (a high-, medium-, or low-specific word) was entered as a fixed effect, and the slope and intercept of the participants as well as the items were entered as random effects (Barr et al., 2013). Since the full model did not converge, we removed the random slope of the item in the model. The *lmer* function from the *lme4* package (Version 1.1-7; Bates et al., 2014) was used to analyze RTs, and the *glmer* function was used to analyze the fixation probability data. The programme was performed in R environment (Version 3.3.2; R Core Team, 2016). The regression coefficients *b*, standard errors *SE*, *Z* values, and *p*-values were reported.

Table 1 shows the mean RTs and accuracies of the button press. The RTs were significantly longer after the participants heard the words in the low-specific condition than in the medium-specific condition ($b = 129.53$,

Table 1. Mean reaction time and accuracy.

	High-specific condition	Medium-specific condition	Low-specific condition
RT (ms)	1691 (58.59)	1631 (55.22)	1759 (62.22)
Accuracy	0.97 (0.01)	0.97 (0.01)	0.94 (0.02)

$SE = 41.34$, $t = 3.134$, $p < .01$), while there was no significant difference between the high- and low-specific conditions ($p = .119$). The accuracy of the button press across all three conditions was high, and was not significantly different across the three conditions ($p > .05$).

The fixation proportions for the targets and the distractors in the high-, medium-, and low-specific conditions from 500 ms before the onset of the spoken target words to 1500 ms after the onset of the target words are shown in Figure 2. An early divergence between the fixations on the targets and the distractors appeared 400 ms after the onset of the spoken words (low-specific condition: $b = 0.225$, $SE = 0.098$, Wald-Z = 2.283, $p < .05$; medium-specific condition: $b = 0.277$, $SE = 0.098$, Wald-Z = 2.833, $p < .01$; high-specific condition: $b = 0.185$, $SE = 0.100$, Wald-Z = 1.845, $p = .07$), after which the fixation probabilities of the targets across all three conditions were significantly higher than those of the distractors (for all conditions, $p < .001$). These results indicate that under all three conditions, the concepts named by the spoken words can direct visual attention to the objects in the display related to the spoken words. Moreover, we analysed the differences in fixation proportions across the high-, medium-, and low-specific conditions (Figure 2). The fixations on the target under the low-specific condition began to separate from the other two conditions in the time window of 500–600 ms (medium-specific condition: $p = .07$; high-specific condition: $p = .06$), and the difference reached a significant level of 600–700 ms ($b = 0.194$, $SE = 0.087$, Wald-Z = −2.222, $p < .05$). That is to say, the targets under the low-specific condition received fewer fixations than under the medium- and high-specific conditions but received more fixations than the distractors. It can also be seen from Figure 2 that the rising trend of fixation proportions under the low-specific condition was slower than under the other two conditions. Furthermore, it shows that the fixation proportions under the medium-specific condition also had a divergence point at 1200–1300 ms ($b = -0.140$, $SE = 0.081$, Wald−Z = −2.519, $p < .05$), after which the spoken words in the high-specific condition caused more fixations than the words in the medium-specific condition.

In the experiment, we controlled the duration time of the spoken words to ensure that the word length did not cause any difference across the conditions. However, three items in the medium-specific condition were 1-syllable words, while all the other words were 2-syllable words. A total of 95.2% of the trials were left after we deleted those 3 items. The pattern of the results was basically unchanged. In addition to the syllable difference, we checked the phonemes of the items in the

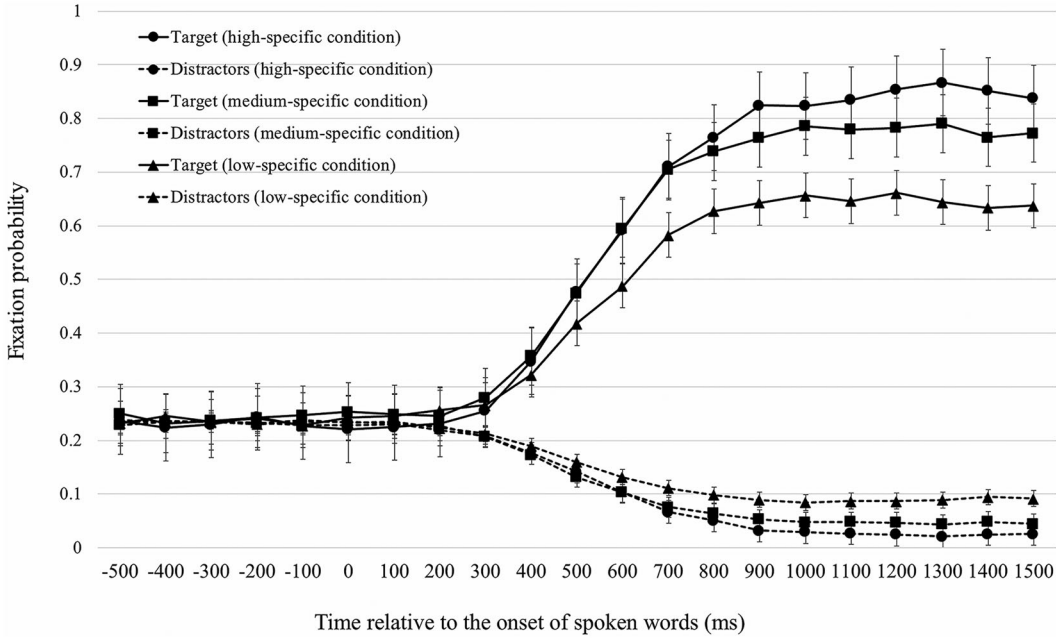

Figure 2. Fixation proportions for the target under the high-, medium-, and low-specific conditions 500 ms before the onset of the spoken target word.

Note: In this figure, the curves of the distractors refer to the mean fixation probability of three distractors in each condition.

study. Only one item was unmatched between the target and distractors. The pattern of results was basically unchanged after we deleted this item. In contrast, syllables are more important than phonemes in terms of the phonology of Chinese words. In the present study, the participants also fixated on the target objects without looking at the distractor that shared one phoneme with the target, further indicating that participants were not sensitive to the phonemes of the Chinese words.

We used the FDR (False Discovery Rate) method proposed by Benjamini and Hochberg (1995) to conduct the *p*-value adjustment. This method is a more powerful adjustment method, which is designed to control the expected proportion of "discoveries" (rejected null hypotheses) that are false (incorrect rejections of the null). The results after the *p*-value adjustment were the same pattern as before the adjustment, except that the difference in fixation probability between the low-specific and medium-specific conditions was marginally significant.

To estimate the difference in the time course of the fixation curve in finer detail, a *growth curve analysis* (GCA) was used (see Mirman et al., 2008). GCA is a type of multilevel regression model that is used for modelling change over time. Mirman et al.' s approach involves two hierarchically related sub-models. The first sub-model (level-1) uses linear regression, assessing the effect of time, including an intercept term, a linear term, a quadratic term, and higher order terms. The use of orthogonal power polynomials makes it possible for the polynomial terms to be independent of each other. The level-2 model describes the level-1 parameters in terms of population averages, fixed effects, and random effects. The intercept term is analogous to the mean fixation proportion, reflecting the average height of the curve. The linear term reflects the overall angle of the curve, that is, the linear slope over the analysis window. The quadratic term corresponds to the degree of curvature, indicating the rise and fall rate around an inflection point. The higher-order terms, such as the cubic and quartic terms, as Mirman et al. stated, may be difficult to interpret and may lack clear cognitive interpretations. In this study, the time course within the analysis window, as plotted in Figure 2, was simple, and the intercept, linear, and quadratic terms were considered likely to be sufficient. Thus, our growth curve models evaluated the effects of these time terms on the fixation proportions, including the critical fixed effect of the conditions and the random effects of the subjects on each time term. Because it takes about 200 ms to programme a saccade (Matin et al., 1993), our analysis window began at 200 ms and extended to 1100 ms when the target fixations reached a plateau.

Table 2 shows the results of this analysis, comparing the targets and distractors under each condition, and the behavioural data and model fits were plotted in Figure 3. The distractor was used as a baseline (via dummy coding). The GCA model reveals a reliable effect of the target on the intercept, reflecting the constant advantage for all types of targets across the window of analysis. In addition, significant effects were found for all the other time terms as well, indicating a steeper slope for the targets and subtle differences in curvature between the curves for the targets and the distractors.

The low-specific condition was used as the baseline for comparing the other two conditions. A reliable effect was found for the intercept, capturing the clear mean difference between the high-, medium-, and low-specific conditions within the time window (see Table 3). Both the high- and medium-specific conditions had higher fixation proportions and steeper slopes than the low-specific condition, as can be observed in Figures 2 and 3. However, differences in the quadratic term were not significant. Moreover, we examined the differences between the high- and medium-specific conditions, but no reliable effect was found for any of the time terms.

General discussion

The study was designed to investigate how eye movements are controlled in the visual world paradigm. The results showed that there were more fixations on the semantically related objects than the distractors when participants heard spoken words from all three hierarchical levels. These results are consistent with previous findings that found visual attention to be sensitive to the semantic relation between visual objects and spoken words (Cooper, 1974; Huettig & Altmann, 2005).

Table 2. GCA results for the fixations of the target and the distractors under each condition.

	High-specific condition				Medium-specific condition				Low-specific condition			
	Est.	*SE*	*t*	*p*	Est.	*SE*	*t*	*p*	Est.	*SE*	*t*	*p*
Intercept	0.46	0.01	35.01	<.001	0.45	0.02	28.89	<.001	0.37	0.02	22.09	<.001
Linear	0.88	0.04	23.19	<.001	0.80	0.05	17.35	<.001	0.66	0.04	16.85	<.001
Quadratic	−0.19	0.02	−7.61	<.001	−0.20	0.03	−6.39	<.001	−0.15	0.03	−4.86	<.001

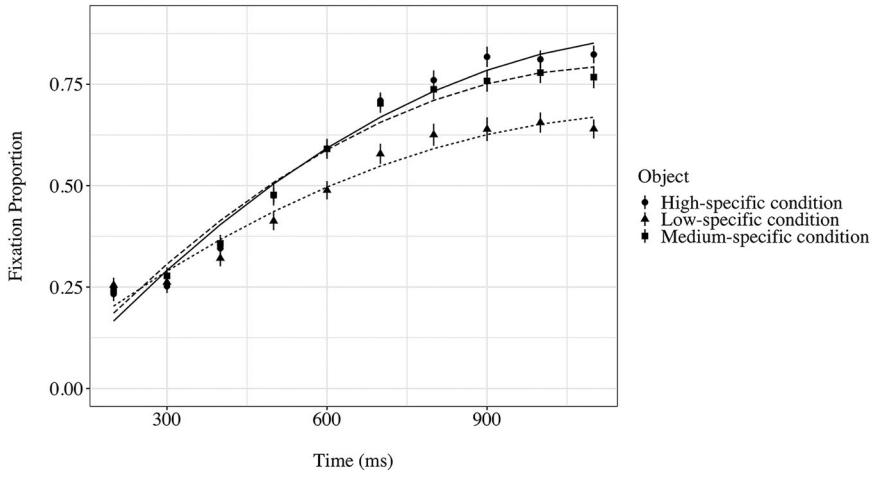

Figure 3. Observed data and model fits for the targets under the High-, Medium-, and Low-specific Conditions.

More importantly, hearing words from different hierarchical levels result in different proportions of eye movements to target objects; words in the high-specific condition cause more fixations to the target object than words in the medium-specific condition, and words in the medium-level condition cause more fixations than words in the low-specific condition. The GCA analysis confirmed these findings.

The current findings are compatible with the *conceptual overlap hypothesis*, showing that different levels of featural overlaps between spoken words and visual objects determine the different probabilities that people will shift their visual attention, such that the greater the featural overlap, the greater the likelihood of a saccade toward the target. Given that the target object in the display (e.g. a picture of a *sparrow*) was precisely the referent of the spoken word (*sparrow*) in the high-specific condition, there was a greater featural overlap between the visual object and the spoken word, leading to more eye movements toward the target objects. The words used in the low-specific condition (*animal*) were highly distinct and lacked specificity, thus presenting the least featural overlap between these spoken words and the visual objects. Thus, participants looked at the target objects the least. The featural overlap is intermediate in the medium-specific condition, leading to an intermediate fixation probability in this condition. In summary, we believe that the degree of conceptual overlap between spoken words and visual objects is the major factor determining the proportion of fixations on a target object.

The finding that featural overlap between visual objects and spoken words plays an important role in driving eye movements is not unexpected. In this study, the visual objects were shown earlier than the spoken words. Before hearing the spoken words, participants might have encoded the semantic information of the visual objects. Therefore, the degree of conceptual overlap between visual objects and spoken words plays an important role in determining where to move the eyes. We can also say that our visual attention is affected by conceptual features that the spoken word can provide. The more features the spoken word provides, the more likely we are to fixate on the target. Our results are consistent with available visual search studies. For example, Maxfield and Zelinsky (2012) proposed that highly specific subordinate representations are best for guidance.

In our results, the early fixations showed no difference in their rising trends between the medium- and high-specific conditions. The GCA shows a steeper slope under the medium- and high-specific conditions than under the low-specific condition. One prominent reason stated for this is that words at the medium hierarchical level are very informative (Murphy & Brownell, 1985; Rosch et al., 1976), and their semantic features are sufficient to distinguish the targets from the distractors, much as a word from the lowest hierarchical level (*sparrow*). As a result, we can map the medium-specific

Table 3. GCA Results for the Target Fixation under the High-, Medium-, and Low-specific Conditions

	High- & low-specific				Medium- & low-specific				High- & medium-specific			
	Est.	*SE*	*t*	*p*	Est.	*SE*	*t*	*p*	Est.	*SE*	*t*	*p*
Intercept	0.09	0.02	5.72	<.001	0.08	0.02	5.24	<.001	0.01	0.01	1.17	0.242
Linear	0.22	0.04	5.63	<.001	0.14	0.04	3.64	<.001	0.08	0.05	1.74	0.082
Quadratic	−0.04	0.03	−1.51	0.131	−0.05	0.03	−1.74	0.081	0.01	0.03	0.45	0.653

word to the target object as fast as the word in the high-specific condition. Then, we can continue to exact more features or information from visual objects and spoken words. The late emergence of discrepancies between the high-specific and medium-level conditions is because more detailed features are beginning to play a part in the fixation probabilities over time.

The results of the current study do not support the *uncertainty reduction hypothesis*, which proposes that the higher uncertainty of the spoken word will cause a higher fixation probability for the visual object. According to this hypothesis, listeners need to look at a corresponding visual object more if the uncertainty of the spoken word is high, so that they can efficiently comprehend spoken language. Thus, we expected higher fixation probabilities when the spoken words had a higher uncertainty or were at a higher abstraction level. Apparently, this is not the case. We found a slower rising trend and a lower probability of fixations toward the target when the spoken word was more uncertain in the low-specific condition.

In conclusion, using the visual world paradigm, we found that the degree of conceptual overlap between spoken words and visual objects affects the probability of people looking at the target object. Participants looked at the target objects more often when there was a greater featural overlap between the visual objects and the spoken words. These results are important for understanding the mechanisms of eye movement control in the interactions between spoken language processing and visual processing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was jointly funded by the National Natural Science Foundation of China (NSFC) and the German Research Foundation (DFG) in Project Crossmodal Learning, NSFC 61621136008/DFG TRR-169. This research was also supported by grants from the National Natural Science Foundation of China [grant number 31970992].

Data availability statement

The data and materials that support the findings of this study are openly available at https://1drv.ms/u/s!Aq3NVEQIAHB-3h2YuJMx_sUCML4z?e=ZEeCpn.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*(38), 419–439. https://doi.org/10.1006/jmla.1997.2558

Altmann, G. T. M., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language*, *57*(4), 502–518. https://doi.org/10.1016/j.jml.2006.12.004

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4* (Version 1.1-7) [Computer software]. http://cran.r-project.org/package=lme4

Battig, W. F., & Montague, W. E. (1969). Category norms of verbal items in 56 categories A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology*, *80*(3, Pt.2), 1–46. https://doi.org/10.1037/h0027577

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, *6*(1), 84–107. https://doi.org/10.1016/0010-0285(74)90005-X

Huettig, F., & Altmann, G. T. M. (2005). Word meaning and the control of eye fixation: Semantic competitor effects and the visual world paradigm. *Cognition*, *96*(1), B23–B32. https://doi.org/10.1016/j.cognition.2004.10.003

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformed or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446. https://doi.org/10.1016/j.jml.2007.11.007

Matin, E., Shao, K. C., & Boff, K. R. (1993). Saccadic overhead: Information-processing time with and without saccades. *Perception & Psychophysics*, *53*(4), 372–380. https://doi.org/10.3758/BF03206780

Maxfield, J. T., & Zelinsky, G. J. (2012). Searching through the hierarchy: How level of target categorization affects visual search. *Visual Cognition*, *20*(10), 1153–1163. https://doi.org/10.1080/13506285.2012.735718

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, *59*(4), 475–494. https://doi.org/10.1016/j.jml.2007.11.006

Murphy, G. L., & Brownell, H. H. (1985). Category differentiation in object recognition. Typicality constraints on the basic category advantage. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(1), 70–84. https://doi.org/10.1037/0278-7393.11.1.70

Myung, J., Blumstein, S. E., & Sedivy, J. C. (2006). Playing on the typewriter, typing on the piano: Manipulation knowledge of

objects. *Cognition*, *98*(3), 223–243. https://doi.org/10.1016/j.cognition.2004.11.010

R. Core Team. (2016). *R: A language and environment for statistical computing* (Version 3.3.2). R Foundation for Statistical Computing. http://www.R-project.org/

Rodrigo, D. B. (2019). *lum_fun for Luminance control of color images* (Version 1) [Computer software]. https://osf.io/auzjy/files/

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439. https://doi.org/10.1016/0010-0285(76)90013-X

Tanenhaus, M. K., Magnuson, J. S., Dahan, D., & Chambers, C. (2000). Eye movements and lexical access in spoken-language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing. *Journal of Psycholinguistic Research*, *29*(6), 557–580. https://doi.org/10.1023/A:1026464108329

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634. https://doi.org/10.1126/science.7777863

Yee, E., & Sedivy, J. (2001). *Using eye movements to track the spread of semantic activation during spoken word recognition*. 13th Annual CUNY Sentence Processing Conference, Philadelphia, March 15th-17th. [Paper presentation].