**Recovery from misinterpretations during online sentence processing**

Lena M Blott1a, Jennifer M Rodd1, Fernanda Ferreira2, and Jane E Warren1

1 Division of Psychology and Language Sciences, University College London, UK

2 Department of Psychology, University of California, Davis, CA, USA

**Supplementary Methods**

**Piloting of ambiguous sentence stimuli**

As part of an exercise to create a large database of ambiguous sentence stimuli for use in multiple experiments, an initial pool of sentence stimuli was constructed around 178 ambiguous words. Ambiguous words (1-3 syllables in length) were selected from published papers reporting meaning dominance for isolated ambiguous words, and included balanced ambiguous words, for which the two most common word meanings were likely to be of approximately equal frequency, and biased ambiguous words, for which one meaning was dominant (Duffy *et al*, 1988; Gorfein *et al.*, 1982; Nelson *et al.*, 1980; Rayner & Frazier, 1989; Rodd *et al.*, 2005; Rodd *et al.*, 2012; Twilley *et al.*, 1994; Vitello *et al.*, 2014; Vuong & Martin, 2011; Wollen *et al.*, 1980). Ambiguous words with a large number of possible meanings were excluded, as were biased ambiguous words with very low-frequency subordinate meanings. A sentence frame (11-15 words; 13-19 syllables) was constructed for each ambiguous word. The ambiguous keyword, which always functioned as a noun, was located near the middle of the sentence with an interval of between 4 and 8 words between the keyword and the final word in the sentence. The neutral sentence opening was consistent with both the subordinate and dominant meanings of the ambiguous word. In the sentences of interest for the present study, the final word of the sentence constrained interpretation towards the subordinate meaning (e.g., “The boys watched the CRANE by the river and saw that it was injured.”). Stimuli for the present eye-tracking experiment were drawn from this pool of piloted subordinate-constrained sentence stimuli. (Additional sentences were also created using the same sentence frames, but with a terminal word that was consistent with either the dominant or subordinate meaning of the ambiguous word; these sentences underwent the same piloting process as the subordinate-constrained sentences but were not used in the eye-tracking study).

Stimuli were recorded by a female speaker of British English. For the purposes of behavioural piloting, the same speaker also recorded the ambiguous words as single-word stimuli, as well as truncated versions of the ambiguous sentences in which the final word of the sentence was omitted. Three phases of behavioural piloting of ambiguous stimuli were conducted in different groups of neurotypical native British English speakers who were predominantly UCL undergraduates, and who did not take part in the eye-tracking experiment;

1. Isolated single-word dominance testing was conducted in one group of neurotypical subjects (N=40) using a two-part computer-based paradigm involving a modified version of the standard word association method (e.g., Twilley et al., 1994), in which participants first generated an associate for each ambiguous word, and then indicated which meaning they had in mind when generating this associate (Vitello *et al.*, 2014). Subjects were not informed prior to the pilot experiment that the words were ambiguous and had multiple possible meanings. The first part of the pilot experiment was a word association task, in which subjects listened to a series of spoken ambiguous words via headphones. After hearing each item, subjects typed the word into the computer (to check that they had heard it correctly), and were then asked to type in a word that was associated with the meaning of the word they had heard. Responses were not time-limited, but subjects were instructed to respond as quickly as possible with the first word that came to mind. The second part of the pilot experiment was a definition selection test, in which subjects heard each of the ambiguous words a second time; each spoken word was followed by visual presentation of their earlier word association response for that word, and a set of two or three written definitions of the ambiguous word. Definitions were brief (1-6 word) descriptions. Two definitions were provided when an ambiguous word had only two potential meanings; where more than two meanings were possible, the three most common meanings were presented. Subjects were asked to select the meaning that they had attributed to the ambiguous word at the time they made their earlier word association response. Responses were indicated by a button press, and were not time-limited. Definition response options were presented in alphabetical order. A final response option, “other”, was also presented for each ambiguous word; if none of the definitions given matched a subject’s preferred meaning, subjects were instructed to select this option, and then manually enter their own definition of the ambiguous word. The first part of the pilot experiment allowed subjects to indicate the meaning of the ambiguous word without being primed to possible meanings, and whilst unaware of the possibility of multiple meanings. The second part of the pilot experiment provided corroboration of subjects’ preferred meanings, and provided clarity in cases where the response generated during the word association task was non-specific or ambiguous.

In order to regulate the length of the experiment, the full set of ambiguous words was divided into 3 subsets containing equal numbers of words and each subject heard the words from 2 of these subsets (i.e. 2/3 of the full stimulus set). Presentation of the subsets was counterbalanced across subjects, and stimuli were presented in a randomised order. The meaning attributed to each ambiguous word was determined by jointly evaluating responses entered for the original word, word association task, and definition selection task. Responses were discarded if the subject’s responses indicated that they had misheard the ambiguous word, if they failed to enter a word association response, or if they indicated that they did not know the meaning of the ambiguous word. Responses were included in dominance calculations if word association and definition selection response were congruent. If these responses were not congruent, responses were excluded unless the meaning indicated by the word association response was clear and unequivocal. Datasets from 3 subjects were judged to be unreliable due to frequent uninterpretable or missing word association responses, and were excluded from the study. After taking into account missing data, dominance scores for each ambiguous word were based on a mean of 24.2 responses per word (range 15-25). Single-word dominance scores for each of the meanings of an ambiguous word were calculated as a proportion of all valid responses for that word.

2. As single-word dominance norms have been found to be inconsistent (e.g. Rice, Beekhuizen, Dubrovsky, Stevenson, & Armstrong, 2019), pilot testing of ambiguous word dominance was also conducted in sentence context. Sentence-embedded dominance testing was conducted via a similar method to the single-word testing using truncated sentence frame stimuli (i.e. omitting the sentence-terminal word) rather than isolated words, in order to enable exclusion of sentence frames that were not neutral (i.e. that markedly altered ambiguous word dominance). Sentence-embedded dominance testing was conducted in a second group of neurotypical native British English speakers (N=40). The same two-part paradigm that was used to assess single-word dominance was used but, instead of single words, subjects heard truncated versions of the ambiguous sentences (sentence frame minus the final word). The procedure followed was similar to the 2-part procedure of the single word dominance testing. First subjects performed a word association task; they heard truncated versions of the sentences, followed by the ambiguous keyword, then typed in the keyword and a related word. Next subjects performed the definition selection task as described above. In order to regulate the length of the experiment, the full set of truncated sentences was divided into 2 subsets containing equal numbers of words and each subject heard the sentences from one of these subsets. The dataset of 1 subject was judged to be unreliable due to frequent uninterpretable or missing responses, and was excluded from the study. After taking into account missing data, dominance scores for each ambiguous word were based on a mean of 19.3 responses per word (range 17-20). Sentence-embedded dominance scores for each of the meanings of an ambiguous word were calculated as a proportion of all valid responses for that word.

On the basis of single word and sentence-based dominance testing, 10 ambiguous words were excluded from the stimulus set. These stimuli were excluded because either: the ambiguous word meaning chosen for subordinate-constrained sentences (on the basis of published data) was not confirmed as the subordinate meaning by our own dominance testing; or keyword dominance did not fit a clear dominant/subordinate pattern (whether biased or balanced), e.g. 3 meanings of roughly equal frequency, or a large number of subordinate meanings such that the dominant meaning also accounted for a relatively low proportion of meanings.

3. Finally, in order to exclude sentences that were perceived as unnatural or difficult to understand, full sentences (i.e. including the terminal word) were tested in a third group of neurotypical subjects (N=24) for comprehensibility, ambiguous keyword meaning preference, and meaning compatibility (ratings of goodness-of-fit of dominant and subordinate meanings within sentence context). *Comprehensibility ratings.* Subjects listened to full-sentence stimuli and were asked to rate ease of sentence comprehension by selecting one of the following response options: ‘Makes sense instantly, no effort required’; ‘Makes sense after a short interval and/or minor effort’; ‘Makes sense after a long interval and/or a lot of effort’; ‘Does not make sense’.

*Ambiguous keyword meaning preference ratings*. After listening to full-sentence stimuli, subjects were presented simultaneously with written definitions of the dominant and subordinate meanings and were asked to rate which meaning(s) fit the sentence context. Meanings were not labeled explicitly as dominant or subordinate but were assigned numbers (i.e. ‘Meaning 1’, ‘Meaning 2’), with random assignation of numbers to subordinate and dominant meanings. Meaning preference selections were made by selecting one of the following response options: ‘Meaning 1 only’, ‘Meaning 2 only’, ‘Both meanings’, ‘Neither meaning’, ‘Don’t know”.

*Meaning compatibility ratings*. After listening to full-sentence stimuli, subjects were presented serially with separate written definitions of the dominant and subordinate meanings (in randomised presentation order), and rated the goodness-of-fit of each meaning with full-sentence context by selecting one of the following response options: ‘Makes perfect sense and sounds natural’, ‘Makes good sense but sounds unnatural or contrived’, ‘Makes some sense but sounds very odd or silly’, ‘Does not make sense at all’.

Each of these measures was obtained for each of the sentences. Due to considerations of pilot experiment length, the set of ambiguous words was split into 4 equal-sized subsets, and each subject was presented with 3 out of the 4 subsets; subset presentation were counterbalanced across subjects and stimulus presentation order was randomised. Subjects were informed prior to the pilot experiment that sentences contained ambiguous words.

Pooling of group data resulted in a maximum of 9 responses per sentence for each of the measures described above. Within each task, data were collated in terms of the total number of subjects selecting a particular response option.

**Selection of eye-tracking stimuli**

Single-word and sentence-embedded dominance data were evaluated alongside comprehensibility and meaning compatibility data in order to select the optimal stimuli for the eye-tracking experiment. Inclusion criteria for stimuli were:

1. *Biased ambiguous keywords*. Stimuli were included only if the sentence-embedded relative meaning frequency for the dominant ambiguous keyword meaning was greater than 0.75 (Sereno, 1995).

2. *Stable ambiguous keyword dominance patterns for isolated vs. sentence-embedded forms*.Stimuli were excluded if the sentence frame altered the dominance pattern of the ambiguous word from biased to balanced, or if the sentence frame reversed the meaning dominance of the ambiguous word so that the subordinate meaning became dominant.

3. *High sentence comprehensibility ratings*. Stimuli were excluded if more than 3 pilot subjects rated a sentence as ‘Does not make sense’.

4. *High levels of subordinate meaning constraint.* Stimuli were excluded if more than 2 pilot subjects judged that the subordinate meaning did not make sense at all, or more than 2 pilot subjects judged that the dominant meaning made perfect sense.

**References**

Duffy, S.A., Morris, R.K., & Rayner, K. (1988). Lexical ambiguity and fixation times in reading. *Journal of Memory and Language, 27*, 429-446.

Gorfein, D.S., Viviani, J.M., & Leddo, J. (1982). Norms as a tool for the study of homography. *Memory & Cognition, 10*, 503-509.

Nelson, D.L., MsEvoy, C.L., Walling, J.R., & Wheeler, J.W. (1980). The University of South Florida homograph norms. *Behavior Research Methods & Instrumentation, 12*, 16-37.

Rayner, K., & Frazier, L. (1989). Selection mechanisms in reading lexically ambiguous words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 779-790.

Rodd, J.M., Davis, M.H., & Johnsrude, I.S. (2005). The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. *Cerebral Cortex, 15*, 1261-1269.

Rodd, J.M., Johnsrude, I.S., & Davis, M.H. (2012). Dissociating frontotemporal contributions to semantic ambiguity resolution in spoken sentences. *Cerebral Cortex, 22*, 1761-73.

Twilley, L.C., Dixon, P., Taylor, D., & Clark, K. (1994). University of Alberta norms of relative meaning frequency for 566 homographs. *Memory & Cognition, 22*, 111-126.

Vitello, S., Warren J.E., Devlin, J.T., & Rodd, J.M. (2014). Roles of frontal and temporal regions in reinterpreting semantically ambiguous sentences. *Frontiers in Human Neuroscience* [Epub ahead of print 29th July 2014], *8*, 530.

Vuong, L.C., & Martin, R.C. (2011). LIFG-based attentional control and the resolution of lexical ambiguities in sentence context. *Brain & Language, 116*, 22-32.

Wollen, K.A., Cox, S.D., Coahran, M.M., Shea, D.S., & Kirby, R.F. (1980). Frequency of occurrence and concreteness ratings of homograph meanings. *Behavior Research Methods & Instrumentation, 12*, 8-15.