**Materials and Methods**

Localizer experiment

*Experimental design and materials*. The design was adopted from Bemis & Pylkkanen (B&P) (2010) and consisted of four conditions (we excluded the two additional conditions that were included in B&P but not used in any of their analyses). The critical composition two-word condition (COMP\_2W) consisted of two words that could be combined into a phrase: a color adjective followed by a noun (e.g., *red boat*). In the control condition (COMP\_1W), the color adjective was replaced by a length-matched unpronounceable consonant string (e.g., *xkq boat*). The list-task conditions were further included to ensure that the difference between the two COMP conditions is not merely due to the fact that the COMP\_2W condition includes an extra word. In the list-task two-word condition (LIST\_2W), participants were therefore presented with two nouns (e.g., *cup boat*; the nouns were always distinct from each other). The control, LIST\_1W condition, was identical to the COMP\_1W condition in terms of the materials (e.g., *xkq boat*).

In all conditions, participants were first presented with the linguistic stimulus (one word at a time, as described in greater detail below), followed by a drawing of a colored shape. They had to evaluate the relationship between the linguistic stimulus and the image. B&P (2010) describe their design as crossing task type (composition task vs. list task) and materials (two words vs. a consonant string and a word) (see Figure 1 in B&P). However, this description is somewhat misleading because although the task differed between the COMP\_2W and LIST\_2W conditions, it was the same between the COMP\_1W and LIST\_1W conditions. In particular, in the COMP\_2W condition, participants were asked to decide whether the image matched the meaning of the linguistic stimulus (e.g., an image of a red boat following “*red boat*” would be a match, but an image of a blue boat, a red cup, or a blue cup would all be non-matches). In the LIST\_2W condition, participants were asked to decide whether the image matched the meaning of either of the two nouns (e.g., an image of either a cup or a boat following “*cup boat*” would be a match, but an image of a plane would be a non-match; the color of the objects was not relevant in this condition). Finally, in the COMP\_1W and LIST\_1W conditions, participants were asked to decide whether the image matched the meaning of the noun (e.g., an image of a boat following “xkq boat” would be a match, but an image of a cup would be a non-match).

Twenty-five one-syllable nouns (disc, plane, bag, lock, cane, hand, key, shoe, bone, square, bell, boat, bow, car, cross, cup, flag, fork, heart, lamp, leaf, note, star, tree, house) were used as the second word. Six adjectives were used in the COMP\_2W condition (red, blue, pink, black, green, brown), and they were matched for number of letters with the nouns used as the first word in the LIST\_2W condition (cup, boat, lamp, plane, cross, house; this was a subset of the 25 nouns above) and with the unpronounceable consonant strings used in the COMP\_1W and LIST\_1W conditions (xkq, qxsw, mtpv, rjdnw, wvcnz, zbxlv). These materials yielded 150 unique phrases for the COMP\_2W condition, 144 unique sequences for the LIST\_2W condition (eliminating the six sequences with duplicate words like “cup cup”), and 150 unique sequences for the COMP\_1W and LIST\_1W conditions. From these sets, for each participant we sampled a semi-random subset of 100 trials for each of the four conditions, so that each of the 25 nouns was used exactly 4 times in the second position within each condition, and so that the words and consonant strings in the first position were matched in letter length, as noted above.

Twenty-five drawings corresponding to the 25 nouns were used, and each could appear in one of six colors, and in one of three orientations (the orientations were randomly sampled from range of 0-360°), resulting in 25 x 6 x 3 = 450 unique drawings (note that for the image of a disc, the images were identical across the orientations). A random scaling factor between 105 and 115% was also applied for each image. The orientation and scaling manipulations were used to diversify the exact physical appearance of the objects.

Several constraints were used in pairing the linguistic materials with the images, which was done separately for each participant (given that the materials differed across participants as described above). In the COMP\_2W condition, for each set of four trials with a particular noun in the second position, there were two trials where the image matched the linguistic stimulus, and two trials where the image did not match. For the non-match trials (a total of 50 in the COMP\_2W condition), there were three options: the object could mismatch (16 trials), the color could mismatch (16 trials), and both the object and the color could mismatch (18 trials). In the LIST\_2W condition, for each set of four trials with a particular noun in the second position, there was one trial where the image matched the first noun, one trial where the image matched the second noun, and two trials where the image did not match either of the two nouns. Finally, in the COMP\_1W and LIST\_1W conditions, for each set of four trials with a particular noun in the second position, there were two trials where the image matched the noun and two trials where the image did not match the noun. The orientation of the objects and their color were chosen randomly, except for the choice of colors in the COMP\_2W condition.

*Procedure.* In both the MEG and fMRI versions, each trial lasted 4s and was structured as follows: an initial fixation (300 ms), a blank screen (300 ms), the first word/consonant string (300 ms), a blank screen (300 ms), the second word (300 ms), a blank screen (300 ms), and finally, the target image presented until the participant responded (by pressing one of two buttons on a button box), for the maximum duration of 2,200 ms (when the participant responded, the image disappeared and a blank screen was presented for rest of the trial). The words and consonant strings were shown in 40 point black Helvetica font in all capital letters in the center of the screen on a light grey background.

Prior to the experiment, participants were provided with printed instructions (see SI), and the experimenter walked them through the instructions and answered any questions that arose. Participants then did a short practice task that consisted of 3 trials per condition and mimicked the actual experiment in terms of the materials (same for all participants) and trial timing.

*MEG version*: Trials were blocked by task (composition task vs. list task), as in B&P’s original experiment. In particular, the 400 trials were divided into four runs of 100 trials each, and each run included only the COMP trials (COMP\_2W and COMP\_1W) or only the LIST trials (LIST\_2W and LIST\_1W). Two run orders were used across participants: i) COMP LIST LIST COMP, or ii) LIST COMP COMP LIST. For each run, the trial order (i.e., the order of 2W and 1W trials) was chosen randomly without replacement from among 8 possible orders, which were created with the constraint that the same trial type (2W or 1W) does not appear more than 3 times in a row, and so that the same answer (Y or N) does not appear more than 4 times in a row. Each trial was followed by an inter-trial interval that varied in length between 200 ms and 600 ms (in increments of 100 ms). Each inter-trial interval appeared exactly 20 times in each run. In addition, a 2 s period was added to the beginning and end of each run, for a total run duration of 444 s (7 min 24 s). Participants were offered short breaks between runs.

*fMRI version*: The experiment used a blocked design, where trials were grouped into blocks of 5 trials of the same condition (so, 20 blocks per condition). Each block therefore lasted 20 s. The 80 blocks were grouped into five runs, with 16 blocks each (4 blocks per condition per run). Each run included a 12 s fixation block at the beginning and end, and after each set of four blocks, for a total run duration of 380 s (6 min 20 s). Condition order for any given run was chosen randomly from among four possible orders. Participants were offered short breaks between runs.

Critical experiment

*Experimental design and materials*. The experiment consisted of two conditions: LOW conceptual specificity and HIGH conceptual specificity nouns. Ninety-six items (pairs of words) were constructed, with each pair consisting of a low- and a high-specificity noun. These pairs spanned diverse semantic domains including animals (e.g., snake-cobra), artifacts (e.g., utensil-fork), clothes (e.g., hat-beret), family relations (e.g., parent-mother), food (e.g., wine-chardonnay), human parts (e.g., bone-femur), inanimate natural objects (e.g., tree-maple), occupations (e.g., doctor-surgeon), sports and leisure (e.g., boat-kayak). Across the set, the words in the two conditions were matched for lexical frequency (LOW specificity: 1770; HIGH specificity: 1723; paired samples t-test: t(95) = -0.10, p = 0.92; Brysbaert et al., 2012) and length in syllables (LOW specificity: 1.99; HIGH specificity: 2.04; t(95) = 0.46, p = 0.65) and letters (LOW specificity: 6.08; HIGH specificity: 6.09; t(95) = 0.04, p = 0.97).

Two behavioral norming studies were conducted on Amazon.com’s Mechanical Turk to evaluate the perceived conceptual specificity of the target words in order to ensure that the materials are well-suited to address the research question at hand. In Norming Study 1, participants (n=20) were asked to rate the conceptual specificity of each word from extremely specific to extremely broad on a 1-7 scale. All participants saw all 192 words. As expected, high-specificity nouns were rated as much more specific than low-specificity nouns (t(95) = 25.47, p < 0.001). In Norming Study 2, a different group of participants (n=20) were presented with the 96 low-/high-specificity pairs (e.g., snake-cobra) and asked to indicate which word in the pair was higher in conceptual specificity. For each word pair, most participants (between 18 and all 20) selected the word that was intended to be higher in specificity. In fact, only 14 out of the total of 1,920 responses (<1%) across participants did not accord with our judgments.

For the MEG and fMRI studies, the 192 words were divided into two experimental lists following a Latin Square design, such that a list contained only one version of an item, but across the experiment (in both MEG and fMRI) each participant saw the trials in both lists to increase experimental power. Each list was further divided into two subsets that corresponded to runs.

In both conditions, participants were first presented with the target word, followed by a yes/no question about the meaning of the word. The questions were included to encourage deep engagement with the materials. Eight questions were used: “Is it alive?”, “Can you eat or drink it?”, “Can you find it in the wild?”, “Can you wear it?”, “Can you hold it?”, “Did humans create it?”, “Does it make sounds?”, and “Does it float in water?”. Each word could be paired with between 4 and all 8 questions. Some questions did not have a clear answer for a particular target word, so we did not use all the questions with every word (for example, the “Can you hold it?” question could be answered either way with the target word “refrigerator” because refrigerators vary in size). For any given participant, each word was paired with one of the possible questions randomly, with the constraints that i) the yes and no correct answers were balanced within each condition in each run across questions (the within-question balancing was not feasible given the highly biased distributions of yes/no responses across questions for this set of materials), and ii) each question appeared 24 times across the 192 trials (and 6 times in each run).

*Procedure*. Each trial was structured as follows: an initial fixation (250 ms), a blank screen (250 ms), the target word (1,000 ms), a blank screen (the duration varied between MEG and fMRI, as described below), and finally, the question presented until the participant responded (by pressing one of two buttons on a button box), for the maximum duration of 3,000 ms (when the participant responded, the question disappeared and a blank screen was presented for rest of the trial). The words and questions were shown in 40 point black Helvetica font in the center of the screen on a light grey background. The target words were presented in all capital letters (e.g., SNAKE), and the questions used normal capitalization (e.g., Is it alive?).

As noted above, the 192 trials were divided into four runs of 48 trials each (24 LOW specificity trials, and 24 HIGH specificity trials). The order of the four subsets of the materials varied across participants. Further, the order of trials in each run was chosen randomly without replacement from among 8 possible orders, which were created with the constraint that the same trial type (LOW or HIGH) does not appear more than 3 times in a row, and so that the same answer (Y or N) does not appear more than 4 times in a row.

As in the localizer experiment, prior to the experiment, participants were provided with printed instructions (see SI), and the experimenter walked them through the instructions and answered any questions that arose. Participants then did a short practice task that consisted of 4 trials per condition and mimicked the actual experiment in terms of the trial timing. The target words used in the practice trials were not used in the actual experiment, but the questions were drawn from the same set of 8 questions used in the experiment.

. In addition, a 4 sec period was added to the beginning of each run,

*MEG version*: The target word and the question were separated by a blank screen presented for 500 ms. Each trial was followed by an inter-trial interval that varied in length between 200 ms and 700 ms (in increments of 100 ms). Each inter-trial interval appears exactly 8 times in each run. In addition, a 2 sec period was added to the beginning of each run, and a 2.4 sec period was added to the end of each run, for a total run duration of 266 s (4 min 26 s). Participants were offered short breaks between runs.