The role of word-word co-occurrence in word learning

Abdellah Fourtassi (a.fourtassi@ueuromed.org)
The Euro-Mediterranean University of Fes
FesShore Park, Fes, Morocco

Emmanuel Dupoux (emmanuel.dupoux@gmail.com)
LSCP/EHESS/ENS
Paris, France

Abstract

A growing body of research on early word learning suggests that learners gather word-object co-occurrence statistics across learning situations. Here we test a new mechanism whereby learners are also sensitive to word-word co-occurrence statistics. Indeed, we find that participants can infer the likely referent of a novel word based on its co-occurrence with other words, in a way that mimics a machine learning algorithm dubbed 'zero-shot learning'. We suggest that the interaction between referential and distributional regularities can bring robustness to the process of word acquisition.

Keywords: word learning; semantics; cross-situational learning; distributional semantic models; zero-shot learning.

Introduction

How do children learn the meanings of words in their native language? This question has intrigued a lot of scholars studying human language acquisition. Quine (1960) famously noted the difficulty of this process. In fact, every naming situation is ambiguous. For example, if I utter the word *gavagai* and point to a rabbit, you may possibly infer that I mean the rabbit, the rabbit's ear, or its tail or color,...etc. A popular proposal in the language acquisition literature suggests that, even if one naming situation is ambiguous, being exposed to many situations allows the learner to narrow down, over time, the set of possible word-object mappings (e.g., Pinker, 1989). This proposed learning mechanism has come to be called Cross-Situational Learning (hereafter, XSL). Laboratory experiments have shown that humans are cognitively equipped to learn in this way. For example, L. Smith and Yu (2008) presented adults with trials that simulated real world uncertainty: each trial was composed of a set of words and a set of objects, in such a way that no single trial had enough information about the precise mappings. However, after being exposed to many of such trials, participants were eventually able to name the objects with a better-than-chance performance. Many experiments replicated this effect with adults, children and infants (Yu & Smith, 2007; Suanda, Mugwanya, & Namy, 2014; Vlach & Johnson, 2013). Subsequent research tried to characterize the algorithmic underpinnings of XSL. Some experiments suggested that learners accumulate in a parallel fashion all statistical regularities about word-object co-occurrences, and they use them to gradually reduce ambiguity across learning situations (McMurray, Horst, & Samuelson, 2012; Vouloumanos, 2008; Yurovsky, Yu, & Smith, 2013). Other experiments suggested that learners maintain, instead, a single hypothesis about the referent of a given word. New evidence either corroborate this hypothesis or contradict it (Medina, Snedeker, Trueswell, & Gleitman, 2011; Trueswell, Medina, Hafri, & Gleitman, 2013). Yurovsky and Frank (2015) proposed a synthesis of both accounts, whereby the learner's choice to adopt one of the two learning strategies depends on the complexity of the learning situation.

This being said, XSL is unlikely to be the unique mechanism of word learning at work. First, real learning situations are much more ambiguous than typical simulated situations used in laboratory experiments. When subjects are tested in a more realistic learning context, the load on memory increases and, therefore, the ability to make use of the available visual information diminishes (Medina et al., 2011; Yurovsky & Frank, 2015). Second, XSL assumes a perfect covariance between words and their referents. This assumption does not take into account the fact that words –in real situations– are sometimes uttered in the absence of their referents (e.g. when talking about past events, "remember that cat?"). In this experiment, we propose a statistical learning mechanism that purports to complements XSL, through relying on cues from the concomitant linguistic information, and more precisely on word co-occurrence.

Form word co-occurrence to semantic similarity

Typical XSL settings assume that words occur in isolation. In real learning contexts, however, words are embedded in natural speech, and have consistent distributional properties. In particular, semantically similar words tend to co-occur more often than semantically unrelated words. For example, the word "ball" and "play" tend to co-occur more often than "ball" and "eat". This fact is documented in linguistics under the name of the 'distributional hypothesis' (hereafter, DH) (Harris, 1954), and has been popularized by Firth's famous quote "You shall know a word by the company it keeps" (Firth, 1957). The distributional hypothesis is also the basis for distributional semantics, the sub-field of computational linguistics that aims at characterizing words' similarity, based on their distributional properties in large text corpora. Tools from the field of distributional semantics such as Latent Semantic Analysis (Landauer & Dumais, 1997), Topic Models (Blei, Ng, & Jordan, 2003), or more recently Neural Networks (Mikolov, Karafiát, Burget, Cernocký, & Khudanpur, 2010) have proved to be effective in modeling human

word similarity judgement (Landauer, McNamara, Dennis, & Kintsch, 2007; Griffiths, Steyvers, & Tenenbaum, 2007; Fourtassi & Dupoux, 2013; Parviz, Johnson, Johnson, & Brock, 2011).

Zero-shot learning

Models that learn through DH typically require a large corpus, especially if nothing is known about the language. Here, we explore the case where some words are already known and only one word is learned through DH. This corresponds to the so-called 'zero-shot learning' situation.

An interesting example of this situation has been given by Socher, Ganjoo, Manning, and Ng (2013). They built a model that can map a label to a picture even when the label has not been used in training! More precisely, using the CIFAR-10 dataset, the model was first trained to map 8 out of the 10 labels ("automobile", "airplane", "ship", "horse", "bird", "dog", "deer", "frog") in the dataset, to their visual instances. The remaining labels ("cat" and "truck") were omitted and reserved for the zero-shot analysis. Second, they used a distributional semantic model (based on Neural Networks) to obtain vector representations for the entire set of labels (i.e., including "cat" and "track") based on their co-occurrence statistics in a large text corpus (Wikipedia text). When tested on it ability to classify a new picture (a cat or a truck) under either the label of "truck" or "cat", the model performed with a high accuracy, using only the patterns of co-occurrence among labels, and the semantic similarity between the new and old pictures. For example, when presented with the picture of a cat, the model has to classify it as "cat" or "truck". The models makes the link between the picture of the cat and that of a similar picture (e.g. dog), and chooses the label that is more related to the label of this similar picture, i.e., "cat". In fact, "cat" co-occurs more often with "dog" than with, say, "airplane". Therefore the label "cat" is favored over the alternative label (i.e., "truck").

The conditions of zero-shot learning are often met in the context of word acquisition. For instance, this corresponds to the (rather ubiquitous) situation where an unknown word is heard in the absence of its visual referent. Therefore, we suggest that human learners can go about it in a way that mimics the mechanism of zero-shot learning. In the following, we test this hypothesis with adults, following closely the spirit of the model developed by Socher et al. (2013).

Method

The experiment consists of 4 steps:

1. Referential familiarization
2. Learning consolidation
3. Distributional familiarization
4. Semantic generalization

The referential familiarization and consolidation consists in explicitly teaching subjects the association between words in an artificial language and their referents. In the distributional familiarization, participants hear 'sentences' made of words from this artificial language without visual referents; some of these words were familiar (introduced in the referential familiarization), and others were novel words. Crucially, the novel items co-occur consistently with words of the same semantic category. Finally, the semantic generalization phase tests whether subjects can rely on distributional information *alone* to infer the semantic category of the novel words, without any prior informative referential situation. Below is a detailed description of each step of the experimental procedure.

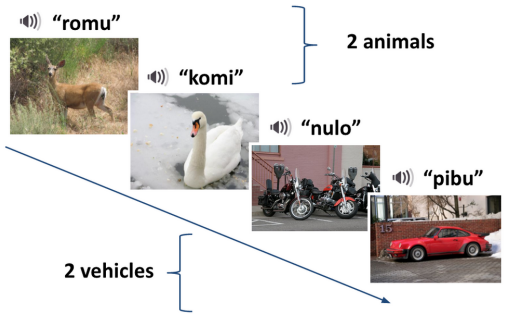

Figure 1: Referential familiarization. Participants are presented with multiple series of word-objects pairings. The objects belong to the category of animals or the category of vehicles.

Step 1: referential familiarization In this phase of the experiment (Figure 1), participants are taught the pairing of 4 words in an artificial language[1] with 4 objects. The objects belong to either the category of vehicles (car, motorcycle) or the category of animals (deer, swan). Participants see a picture of the referent on the screen and hear its label simultaneously. There are 3 trials, each consists of a randomized presentation of the series of 4 pairings.

Step 2: learning consolidation The purpose of this phase is to consolidate and strengthen the participants' knowledge about the 4 word-object pairings (Figure 2). Participants are tested using a Two Alternative Forced Choice paradigm (2AFC). They are presented with a series of trials where they hear a label (*pibu*, *nulo*, *romu* or *komi*) and are shown two objects; one of which is the correct referent, and the other belongs to the other semantic category. Crucially, after they have made a choice, they get a feedback on their answers ("correct"/"wrong"). Participants are presented with 16 questions of this sort, which correspond to the combinatorial possibilities of forming pairs of items from one semantic category with items from the other category (4 cases), in conjunction with the order of the visual presentation of the referents (4×2 cases) and the item being labeled ($4 \times 2 \times 2 = 16$ cases in total).

[1] The audio stimuli were graciously provided by Naomi Feldman

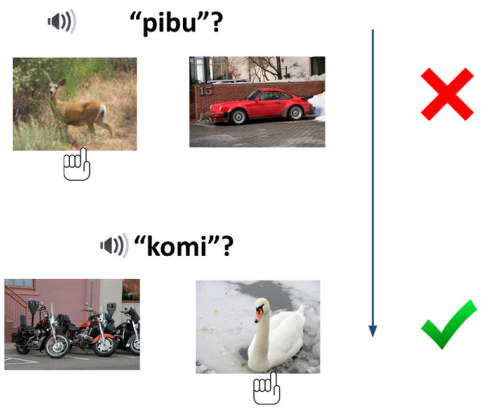

Figure 2: Learning consolidation. Two-Alternative Forced Choice paradigm (2AFC), with feedback.

Step 3: distributional familiarization Distributional familiarization follows the referential training and consolidation. Participants listen to 'sentences' made of words from this artificial language without any visual referent. As explained in Figure 3, each sentence consists of 3 words. Two of which are familar words from one semantic category, i.e., either *romu* and *komi* (animals) or *pibu* and *nulo* (vehicles). The third word is a new artificial word that consistently co-occur with them. The new words are *guta* and *lita*. The way *guta*/*lita* are distributed with either (*romu*, *komi*) or (*pibu*, *nulo*) was counterbalanced across participants so as to avoid different sorts of linguistic and perceptual biases that may arise from the way the stimulus is organized. There is a 750 ms pause between words, and 2500 ms pause between sentences. There are 16 sentences in total, 8 for each semantic context; (*romu*, *komi*) and (*pibu*, *nulo*). Words within sentences are randomized and the semantic context is alternated during the exposure.

Step 4: testing semantic generalization Participants are presented again with a two alternative forced choice. As explained previously in the learning consolidation phase, they hear a label and they are asked to choose between two objects, but here participants do not get feedback on their answers. We are particularly interested in how participants respond in the situation where they hear the novel item (*guta* or *lita*) and are presented with two new objects that represent a new animal (squirrel) and a new vehicle (trolley). Participants have never been shown the referential mapping of the new words, so their answer would reveal whether distributional learning alone had helped them infer semantic knowledge about the word (i.e., the semantic category of the referent). This test phase is composed of 4 questions about the novel labels/objects, varying the visual order of the objects (1×2) and the object being named ($1 \times 2 \times 2 = 4$ cases in total), in addition to 4 selected questions about the familiar words/objects used in the referential training. We eliminated any overlap between questions about novel items and questions about familiar items so as to avoid any form of cross-situational learning during the test phase.

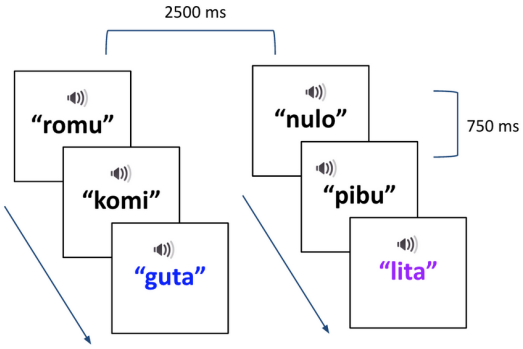

Figure 3: Distributional familiarization. Sequences of words are presented with no visual referents. Two new words ("guta" and "lita") are introduced and co-occur consistently with the words corresponding to one of the two semantic categories ("romu" and "komi" for the category of animals, and "nulo" and "pibu" for the category of vehicles)

Procedure As shown in Figure 4, participants are first trained on the pairing between 4 artificial words and their referents (part 1 and part 2). Then they are exposed to 2 blocks of distributional familiarization (part 3), and they are tested 3 times (part 4): before any exposure to distributional information (baseline) and after the first and the second block of distributional exposure (respectively session 1 and 2).

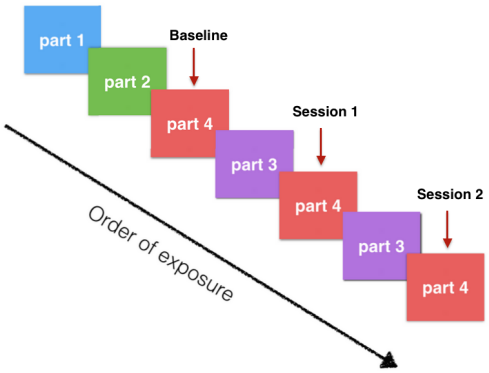

Figure 4: Order of exposure in the experiment. Participants are trained referentially once (part 1 and part 2), distributionally twice (part 3). They are tested in three sessions (part 4): before and after each block of distributional learning

Population and rejection criterion 50 Participants were recruited online through Amazon Mechanical Turk. We included in the analysis participants whose total score on the familiar word-objet questions during the testing phases (i.e., part 4) were above chance level. This is a way to select only subjects who paid attention during the training parts. 2 par-

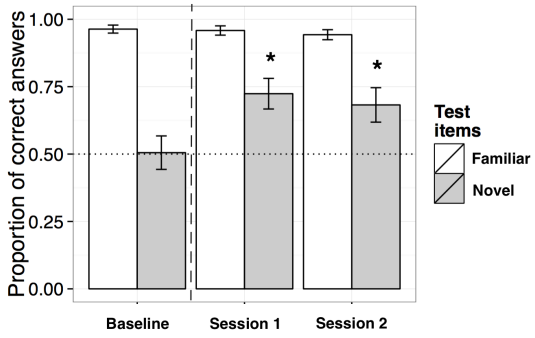

Figure 5: proportion of correct answers for familiar and novel test items, before any distributional exposure (baseline) and after the first and second block of exposure (session 1 and 2)

ticipants was excluded based on this criterion.

Results and Analysis

Figure 5 shows the proportion of correct answers on both familiar and novel items, as a function of the testing session. In the familiar condition, answers were almost perfect in the three sessions (before exposure, after one block, and after two blocks of exposure to part 3). This shows that participants have reliably learned the association between words and their referents during the training phase, and that this learning was not affected by subsequent exposure to distributional information. In the novel condition, and before distributional training (i.e., baseline), subjects were at chance level ($M = 50.5\%$ of correct answers). A one sample t-test comparing the mean against chance (i.e, 50%) gives a $t(47) = 0.083$ with p-value $= 0.93$. The absence of learning is a predictable result since participants had no prior cue about the relevant object mapping. However, after one and two blocks of distributional training, subjects were significantly above chance level. A one sample t-test gives, respectively, for session 1 an average of correct answers $M = 72.4\%$, with $t(47) = 3.94$ ($p < 0.001$), and for session 2, an average of $M = 68.2\%$, with $t(47) = 2.85$ ($p = 0.006$). In order to compare the behaviour of the participants before and after distributional training, we performed a paired t-test. For baseline vs. session 1, there was a significant change, the difference mean is equal to $M = 0.218$, with $t(47) = 2.99$ ($p < 0.01$). Similarly, for baseline vs. session 2, the difference mean is $M = 0.177$, with $t(47) = 2.24$ ($p = 0.029$). However, between session 1 and session 2, the difference mean $M = 0.041$ was not significant, $t(47) = 0.662$, $p = 0.51$. This shows that most of the learning occurred during the first block of distributional exposure. Additional training did not significantly improve learning (if anything, it seems to slightly decrease the average of correct responses).

Discussion

The results show that, when learning the meaning of words, people are sensitive, not only to the co-occurrence of words and objects (as suggested in XSL), but also to co-occurrence statistics between words themselves (as suggested in the DH). More importantly, we showed that these two sensitivities interact in a way that mimics a machine learning mechanism called zero-shot learning. In fact, participants in our experiment were able to guess the semantic category of a novel word whose visual referent was never presented through the semantic properties of the words with which it co-occurred consistently. Participants knew beforehand that they would be introduced to an artificial language and that they would have to learn the meaning of words in this language, but they were not explicitly instructed about the fact that words that co-occur in the same sentences are supposed to have similar meanings. Participants have spontaneously turned to co-occurrence in order to cue semantic similarity, and infer the category of the ambiguous words.

Although we used an artificial language whose 'sentences' fall short, on many aspects, of real speech, this work provides evidence for the *cognitive plausibility* of this learning mechanism, much in the spirit of the statistical learning literature (e.g., L. Smith & Yu, 2008; Saffran, Aslin, & Newport, 1996). If it scales up to real languages, this word-word co-occurrence mechanism would prove crucial in complementing word-object co-occurrence mechanisms. In fact, most word-object co-occurrence learning strategies (e.g. XSL) assume that words covary perfectly with their referents. This assumption is not always correct. For example, when talking about a past event, the conversation may not match the immediate visual context. In contrast, words used in a given conversation, be it about present, past or future events, normally co-occur in a coherent fashion. The learner can rely on this intrinsic property of speech to bring about robustness to the learning process. For example, suppose the learner, while at home, hears a discussion about the last visit to the "zoo". XSL learning, if operating alone, would be confusing. In contrast, if XSL operates in concert with DH, the learner would tend, if in doubt, to link a new word (e.g., "zoo") not to some surrounding object, but to other co-occurring words, which are likely to be zoo-related words (such as "animals", "bird" and "monkey"). Further work is needed to characterize the precise conditions under which learners would rather switch to the word-word co-occurrence cue to infer meaning.

Moreover, the proposed mechanism can help learners develop an early semantic representation for words with a rather abstract meaning. Abstract words (like "eat" and "good") are learned later in development than words with salient concrete referents (such as "ball" and "shoe") (e.g., Bergelson & Swingley, 2013). They are presumably harder to learn because there is no obvious or/and lasting correspondence between the word and the physical environment. Bruni, Tran, and Baroni (2014) developed a model which extends purely word-word co-occurrence learning strategies (such as LSA model) to also encompass co-occurrence with the visual context. They assessed the contribution of textual and visual information in approximating the meaning of abstract vs. con-

crete words. They found that visual information was mostly beneficial in the concrete domain, while it maintained an almost neutral impact on the abstract domain where most learning was based on word-word co-occurrence. Future work will investigate the extent to which this finding squares with psychological behaviour. For instance, an interesting question would be to test whether human learners switch from word-object cue to word-word cue when the potential abstractness of the target word increases.

Finally, during the write-up of this paper, it came to our knowledge that Ouyang, Boroditsky, and Frank (in press) conducted an experiment that shared many similarities with ours. However, it also presented interesting differences both in terms of the experimental setup and the results. Ouyand at al. exposed adult participants to auditory sentences from a MNPQ language. It is an artificial language where sentences take the form of "M and N" or "P and Q". Ms and Ps are used as context words, whereas Ns and Qs are target words. We believe there are two crucial differences between the two experiments. First, the context words (M and P) were composed of a mix of various proportions of real English words or non-words. In our experiment, they were all non-words. Second and more important, Ouyang et al. (in press) followed the spirit of MNPQ's paradigm in keeping constant the order of the words in the sentences, that is, M and P always occurring first in the sentence, and N and Q always occurring last. Our experiment was more faithful to the hypothesis of *bag-of-words*, which is crucial in distributional semantic models: order within a particular semantic context (e.g., a sentence) is irrelevant. It was therefore randomized across trials. Interestingly, although none of the context words we used were known words, we obtained a high learning rate. In contrast, Ouyang et al. (in press) obtained successful learning only when most of the context words were familiar English words. A plausible explanation for this difference is that, in the case of MNPQ language, participants have two possible learning dimensions: learning the positional patterns (what word comes first, and what words comes last) and learning the co-occurrence patterns (what couple of words co-occurred with each other). In fact, it has been shown that when both positional and co-occurrence cues are available, participant tend focus on the first ones (K. Smith, 1966). By using familiar words, Ouyang et al. (in press) showed that participants were more likely to learn co-occurrence patterns, probably through alleviating part of the memory constraint. In our case, the positional patterns was random, which left participants with only one learning dimension (i.e., co-occurrence pattern).

To conclude, this experiment provided a cognitive proof of principle to the zero-shot learning mechanism, according to which a (early) semantic knowledge can be learned through sensitivity to word co-occurrence in speech. Future work will focus on exploring properties of this new learning and how it interacts with cross-situational learning.

Acknowledgments

This work was supported by the European Research Council (ERC-2011-AdG-295810 BOOTPHON), the Agence Nationale pour la Recherche (ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL*), the École des Neurosciences de Paris Ile-de-France, the Fondation de France, the Region Ile de France (DIM Cerveau et Pensée), and an AWS in Education Research Grant award.

References

Bergelson, E., & Swingley, D. (2013). The acquisition of abstract words by young infants. *Cognition*, *127*.

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Bruni, E., Tran, N., & Baroni, M. (2014). Multimodal Distributional Semantics. *Journal of Artificial Intelligence Research*, *49*.

Firth, J. R. (1957). A synopsis of linguistic theory 19301955. In *Studies in linguistic analysis.* Oxford, Blackwell.

Fourtassi, A., & Dupoux, E. (2013). A corpus-based evaluation method for distributional semantic models. In *Proceedings of ACL.*

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211-244.

Harris, Z. (1954). Distributional structure. *Word*, *10*(23), 146-162.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, *104*(2), 211-240.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2007). Handbook of latent semantic analysis. Mahwah, NJ: Erlbaum.

McMurray, B., Horst, J. S., & Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review*, *119*.

Medina, T., Snedeker, J., Trueswell, J., & Gleitman, L. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, *108*(22), 9014.

Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In *Proceedings of INTERSPEECH.*

Ouyang, L., Boroditsky, L., & Frank, M. C. (in press). Semantic coherence facilitates distributional learning of word meaning. *Cognitive Science*.

Parviz, M., Johnson, M., Johnson, B., & Brock, J. (2011). Using language models and latent semantic analysis to characterise the n400m neural response. In *Proceedings of the Australasian Language Technology Association Workshop.*

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT press.

Quine, W. (1960). *Word and object*. The MIT Press.

Saffran, J. R., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926.

Smith, K. (1966). Grammatical intrusions in the recall of structured letter pairs: mediated transfer or position learning? *Journal of Experimental Psychology*, *72*, 580–588.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568.

Socher, R., Ganjoo, M., Manning, C. D., & Ng, A. Y. (2013). Zero-Shot Learning Through Cross-Modal Transfer. In *Proceedings of Conference on Neural Information Processing Systems (NIPS)*.

Suanda, S. H., Mugwanya, N., & Namy, L. L. (2014). Cross-situational statistical word learning in young children. *Journal of Experimental Child Psychology*, *126*.

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational learning. *Cognitive Psychology*, *66*.

Vlach, H. A., & Johnson, S. P. (2013). Memory constraints on infants' cross-situational statistical learning. *Cognition*, *127*.

Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition*, *107*(2), 729–742.

Yu, C., & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*(5), 414–420.

Yurovsky, D., & Frank, M. C. (2015). An Integrative Account of Constraints on Cross-Situational Learning. *Cognition*, *145*.

Yurovsky, D., Yu, C., & Smith, L. B. (2013). Competitive processes in cross-situational word learning. *Cognitive Science*, *37*.