Statistical Learning of Conjunctive Probabilities

Di Mo (di.mo@mail.utoronto.ca)
Department of Psychology
University of Toronto

Blair C. Armstrong (blair.armstrong@utoronro.ca)
Department of Psychology
University of Toronto

Abstract

Most statistical learning studies focus on the learning of transitional probabilities between adjacent elements in a sequence, however, other statistical regularities may underpin different aspects of processing language and regularities in other domains. Here, we investigate how conjunctive statistical regularities (of the form A and B together predict C) can be learned, and how this learning is impacted by similarity in representations analogous to that in unambiguous words, homonyms with multiple unrelated meanings, and polysemes with multiple related meanings. We observed that provided the stimulus structure is relatively simple, participants are readily able to learn conjunctive probabilities and display sensitivity to relatedness among representations. These results open new theoretical possibilities for exploring the domain-generality of how the learning and processing systems merge conjunctive information in simple laboratory tasks and in natural language.

Keywords: Statistical Learning; Lexical Ambiguity; Transitional Probability; Conjunctive Probability

Introduction

Statistical learning has been proposed as a powerful mechanism for how individuals learn regularities across time and space. Foundational work by Saffran, Newport, and Aslin (1996) first established human sensitivity to transitional probabilities (TPs) in identifying word boundaries in streams of auditory syllables. Most research on this subject to date has focused on variations of TPs such as non-adjacent dependencies (Gómez, 2002) and visual co-occurrences across scenes (Fiser & Aslin, 2001), illustrating a range of applications for statistical learning. While fundamental, the various forms of TPs do not account for all types of statistical regularities that must be learnt to explain other types of behaviours. For example, learning something akin to a conjunctive probability (CP) may be important in explaining how individuals learn to disambiguate the meanings of semantically ambiguous words in natural language. To illustrate, the word BAT can refer to either an animal or to sporting equipment, and the correct meaning of this word is extracted by integrating the constraints on overall meaning offered by BAT with the broader context (e.g., a discussion about baseball).

The present work sought to investigate several major issues that relate to learning CPs, as they might relate to natural language statistics such as those relevant to word meaning disambiguation. The first was how different elements in a stream could be more or less constraining on the expected outcome of a conjunction. For example, in natural language, knowing that the topic of conversation is “SPORTS” provides only vague constraint on what particular meaning should be evoked in a sentence. This knowledge therefore provides only low constraint (high entropy) in determining which particular meaning should be evoked (e.g., the discussion could relate to hockey, baseball, etc.). In contrast, the word “BAT” provides relatively high constraint (low entropy) on what meaning should be evoked (it should relate either to “baseball” or to “flying mammal”). Furthermore, critical to present purposes, only by combining both of these elements can a context-specific interpretation of a word be evoked. Using this analogy to words (which are low entropy), contexts (which are high entropy), and context-specific meanings (which are fully determined by the combination of the previous two elements) we examined how low- and high-entropy items combined to predict an upcoming element. In a related vein, we also examined how the order in which low- versus high-entropy information is presented shaped performance. How is the process of computing CPs impacted by having more versus less constraint early in processing?

Additionally, unlike typical statistical learning research which employs highly and equally distinct elements during learning, we also explored how representational similarity could shape performance in computing a CP and relate to word disambiguation processes. In the case of natural language, the semantic ambiguity continuum can be broken down into three main subdivisions: (1) unambiguous words like CHALK which evoke effectively the same meaning in different contexts. That is, the word itself predicts the meaning with 100% accuracy, the context does not provide any additional unique information. (2) homonyms such as BANK which evoke completely distinct meanings in different contexts. That is, the word narrows the meaning down to two completely distinct interpretations, but context is necessary to select among those representations. (3) polysemes such as CHICKEN, which evoke related representations (in this example, the animal or its meat) in distinct contexts. That is, the word alone may predict the majority of the evoked representation, but context is needed to select exactly the right interpretation.

With these aims in mind, we developed a variant of a standard self-paced statistical learning paradigm that allowed us to contrast standard TP learning with the learning of CPs between low-entropy items (analogous to words) and high-entropy items (analogous to contexts) in predicting a third item (analogous to context-sensitive meaning). We also employed representations that varied in their similarity to one another to assess the impact of meaning relatedness on learning. Performance was assessed using a combination of online and offline measures of learning. In so doing, we aimed to contribute to knowledge of how a broader range of statistics such as conjunctive probabilities can be incorporated into general theories of statistical learning. We also aimed to connect this work with important statistical properties that are at the heart of other areas of cognition such as semantic ambiguity resolution. If successful, this work could open new possibilities for how artificial language learning experiments using statistical learning paradigms could complement existing studies of semantic ambiguity in natural language, for example, by allowing the development of well controlled artificial languages that avoid the complex confounds in natural language stimuli used to study semantic ambiguity (Armstrong & Plaut, 2016).

Experiment 1

The first experiment served as a baseline for evaluating how the learning of standard triplet structures with perfect predictability (TPs of 1) across successive items takes place using our specific experimental procedure. We then use these results as a platform for understanding the impact of ambiguity on processing in subsequent experiments using variations of the same basic design but changing the probability structure between elements.

Methods

Participants A total of 60 participants (16 male; mean age=20) completed the experiment. All participants were undergraduate students from the University of Toronto participant pool and were compensated with course credit. All completed an informed consent and debriefing procedure.

Materials A total of 48 images of unusual objects (hereafter, symbols) were the targets for learning in the experiment. These symbols were selected so as to not have clear verbalisable labels, and therefore encourage learning of the statistics between the visual representations of each element. These symbols were used to create sequences containing two single-symbol elements and one four-symbol complex element. Eight such simple-simple-complex sequences with unique elements were randomly generated for each participant. The use of varying complexity across visual elements (one symbol vs. four symbols) allows us to assess the impact of visual complexity per se, and also enables rich variation in the statistical structure of the relationship between elements and symbols in the subsequent experiments.

Procedure The experiment was administered on desktop computers using PsychoPy (v1.85.4).

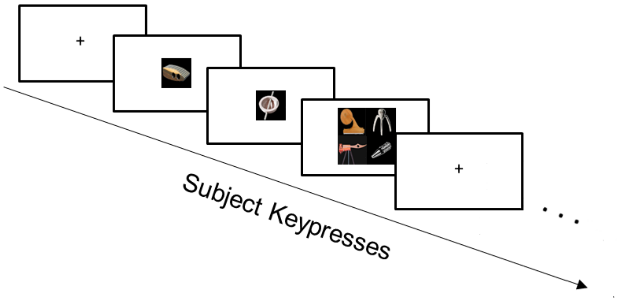

Figure 1: Familiarisation

Familiarisation/On-line Learning Participants were exposed to 30 randomised sweeps through the eight sequences and were instructed to pay attention to the order of the elements. A fixation cross was presented between sequences to focus learning on the relationships between elements (see Figure 1). The task was self-paced and participants advanced through the elements by pressing the space key. The time spent on each element was recorded. On average, the familiarisation task took approximately 20 minutes to complete.

Off-line Tests Two offline tasks were used to assess learning. The first was a sequence completion task, in which participants had to complete a missing element in a sequence. Participants selected from among four choices for completing the first and last element, and two choices for completing the middle element. This corresponded to later experiments where one of the first two elements had only two valid possibilities. The presented choices all came from the same position in a sequence, sampled from among the different sequences (e.g., the choices were always taken from position 1 when completing a missing element from position 1). Eight questions each were asked about the first two elements and 12 questions were asked about the third element. The four extra questions about the third element in this experiment were only included in order to match the number of questions used in subsequent experiments regarding CPs (as described later). Test questions were blocked by order of position in the sequence.

The second task had participants choose from among four sequences which was the most familiar. One of these sequences was actually seen during familiarisation, the others were made-up sequences that mixed elements from different sequences while preserving position in a sequence (e.g., a sequence would be made up of an element selected at random from all elements in position 1 across sequences, an element selected at random from position 2 across sequences, etc.). Sequences were presented one element at a time at a fixed rate of one element per second. Six questions were asked: two for coarse-grained distinction, where all non-target sequences comprised entirely unfamiliar combinations of elements; four for fine-grained distinction, which included a distractor item containing two elements from one sequence combined with one element from another sequence. Again, number of questions were matched to those of subsequent experiments on conjunctive probability. While only one element is needed to predict a sequence in TP, subsequent conjunctive probability experiments will require looking at two elements together to predict the third.

Results

Due to space constraints, we report only the differences that were significant at p¡0.05. Error bars in graphs denote standard error. For this experiment, it is expected that there is a speed up between Elements 1 and 2 as the first element is unpredictable while the second element is perfectly predictable from the first element. It is also expected that the last element requires more processing effort due to higher visual complexity.

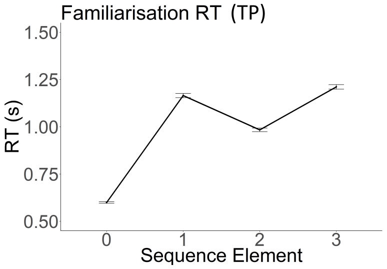

Figure 2: Element 0=fixation; 1-3=sequence

Familiarisation The average reaction time (RT) during online familiarization to the sequences is presented in Figure 2. We used mixed-effect linear models with random intercepts for participants to test for differences in RT across sequence elements (positions) 1-3. Participants sped up between Elements 1 and 2 but slowed down between Elements 2 and 3 such that Element 3 took significantly longer time to process than Element 1.

Offline Test One sample t-tests showed participants had learnt all three elements in the sequence above chance performance in the sequence completion task, as reflected in their accuracy in questions regarding each element. Note that due to the aforementioned difference in number of options at test (but not in training), chance is 0.25 for Elements 1 and 3 and 0.5 for Element 2 (Figure 3). To compare relative learning across the sequence, accuracy was modeled with items and participants as random intercept. Significant differences were found among all three elements, with the highest accuracy at Element 2 and lowest accuracy at Element 3. However, due to difference in chance level, only Elements 2 and 3 can be directly compared. Above-chance performance was also found for all elements in sequence selection for familiarity. There were no significant differences between coarse-grained and fine-grained test items.

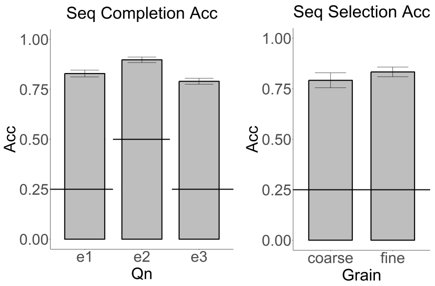

Figure 3: Solid lines mark chance performance.

Discussion

Experiment 1 showed that participants were sensitive to the statistical properties associated with each sequence element in the online learning measure. Participants displayed good performance (7̃5% correct) on all the elements in the offline test. Both of these results are consistent with a similar prior study by (Siegelman, Bogaerts, Kronenfeld, & Frost, 2018). In contrast to that experiment, however, participants exhibited overall slower responses for the final element in the sequence, which we attribute to the increased visual complexity of that item. These results provide an important measure of baseline performance in the task to evaluate the impact of CP learning in the following experiments.

Experiment 2

Experiment 2 used the same overall procedure but different statistical relationships between elements to probe how CPs, as well as different levels of ambiguity, influence behaviour. As in the case of natural language, disambiguating information can be precede or follow an ambiguous word. Hence, two sub-experiments were run, in which the order of the first two elements were interchanged so that the first element either provided high constraint (low entropy, Expt 2a) or low constraint (high entropy, Expt 2b) for predicting the final element, which was the same in both experiments. The experiments thus evaluated the impact of conjunctive probability learning and on the order of the more constraining versus less constraining elements on learning. If people integrate information in a manner analogous to CPs, it is expected that they would show slowdown according to ambiguity type, as illustrated by overlaps in sequence elements, over and beyond slowdown caused by visual complexity. We also expect differences as a result of informativeness of different elements. However, whether more informative elements will be faster to process due to the time to hone in on a specific interpretation or slower due to the number of competing predictions is not clear.

Methods

Participants. A separate sample of 60 undergraduate participants who have not participated in other experiments were recruited for each of the experiments (2a: 15 male; mean age=19; 2b: 22 male; mean age=19).

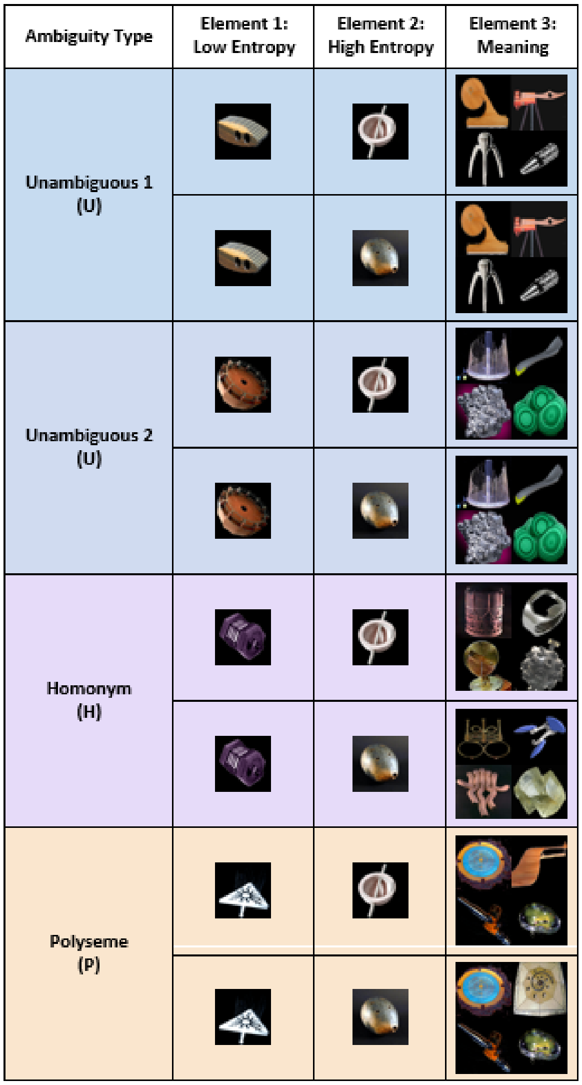
Ambiguity Type
Element 1: Low Entropy
Element 2: High Entropy
Element 3: Meaning
Unambiguous 1 (U)
Unambiguous 2 (U)
Homonym (H)
Polyseme (P)

Figure 4: Example sequences depicting ambiguity types.

Materials. The same elements used in Experiment 1 were re-arranged to reflect different statistical relationships between the items, both in terms of how well each of the first two elements predicted the last element, and in terms of how distinct the last element is relative to its counterpart. These sequences were structured to represent three levels of ambiguity in how the low-entropy (word) representation merged with the high-entropy (context) representation. Across two contexts, Element 3 in an unambiguous sequence was identical, Element 3 in a polyseme sequence overlapped by 25% (one symbol), and Element 3 in a homonym sequence was distinct (see Figure 4). Single-symbol elements were used for the low- and high-entropy elements (words and contexts), whereas four-symbol elements were used to denote ”meanings”, so as to enable studying the effects of representational overlap. Symbols forming each element were randomized across participants.

Procedure The procedure was identical to that in Experiment 1, except the items were re-arranged to have the conjunctive probability structure outlined above and illustrated in Figure 4 for Experiment 2a (in Expt 2b, the position of the low-entropy and high-entropy items were swapped). The sequence completion task now contained eight coarse-grain and four fine-grain questions regarding Element 3 (meaning) instead of 12 questions of equal difficulty. In this experiment, fine-grained questions for both off-line tests refer to items where the choices given contain both options corresponding to the two possibly correct third elements, depending on context. Because the unambiguous sequences evoke the same meaning (Element 3) regardless of context (Expt 2a: Element 2; Expt 2b: Element 1), tests relating to the second item (first item in Expt 2b) were omitted since both context items were valid responses. This left six sequence familiarity items. Having more trials for one offline task type was due to our aims of efficiently extracting the learning of coarse- and fine-grained information.

Results

The analytical procedures were the mostly the same as Experiment 1, only now we collapsed across performance of the same ambiguity type and applied the linear model within each type.

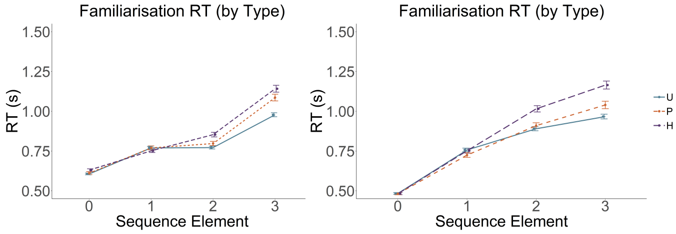

Figure 5: *Experiment 2a.* Element 0=Fixation; 1=Low Entropy; 2=High Entropy; 3=Meaning (left)
Experiment 2b. Element 0=Fixation; 1=High Entropy; 2=Low Entropy; 3=Meaning (right)
U = Unambiguous; P = Polyseme; H = Homonym

Familiarisation *Experiment 2a.* Figure 5 plots the results from familiarization for Experiment 2 and 2b. In Experiment 2a, RT for homonym sequences showed increase across all consecutive elements while polyseme sequences and unambiguous sequences showed slowdown only from Element 2 to Element 3. At the second position (high-entropy), the linear model for RT against ambiguity type showed homonym sequences to be sig-

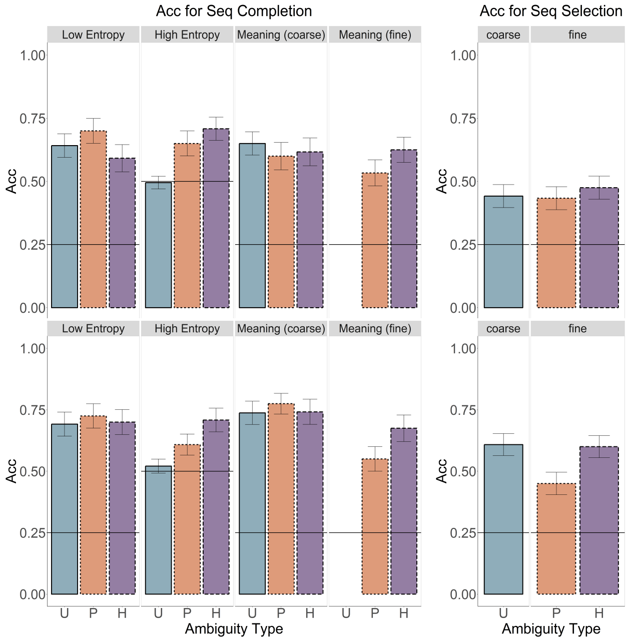

Figure 6: Experiment 2a (top) and 2b (bottom) offline tests. Horizontal lines denote chance.

nificantly different from unambiguous and polyseme sequences, which were comparable to each other. All three conditions were significantly different at meaning output (Element 3), with fastest performance for unambiguous items and slowest performance for homonym items.

Experiment 2b. In every ambiguity type, RT increased between consecutive elements. Similar to experiment 2a, there was no difference between ambiguity types in the first position, but at the second position (now low-entropy), divergence began where homonym sequences showed significant difference from unambiguous and polyseme sequences and by the third position (meaning output) all three ambiguity types were significantly different from each other.

Offline Test *Experiment 2a.* Figure 6 shows the offline task accuracy for experiment 2a and 2b. One sample t-test showed above-chance performance for all questions, indicating learning. A linear mixed effect model showed higher performance for polysemes than homonyms in the second position (low entropy) but no other differences between ambiguity types. Participants also performed above chance for the selection of familiar sequence. Performance did not differ by ambiguity.

Experiment 2b. All question types had above-chance accuracy. Fine-grained meaning (Element 3) in homonym was significantly more accurate than polyseme. There was comparable performance between ambiguity types for other elements. In the sequence selection task, performance were significantly above chance for all ambiguity types. Linear mixed model showed significant differences between polyseme and homonym sequences. Furthermore polyseme sequences were significantly more affected by the presence of context-inappropriate foils than homonym sequences.

Discussion

Experiment 2 showed an increased slowdown starting at the integration of contextual element according to the increased overlap in interpretations across contexts. In contrast to the unambiguous items and to the results obtained in Experiment 1, participants were slower to respond in the online task when learning CPs in ambiguous sequences. The amount of slowdown in the online task showed that these effects were modulated both by the amount of overlap in the meaning representations, and whether the more informative (lower entropy) item was presented earlier or later in the sequence. The lack of differentiation at Element 1 suggested that only with two elements was there enough information to integrate in order to predict the third element based on CPs. This is different from words in context-free tasks (Armstrong & Plaut, 2016) and tasks with contextual constraints for natural language (Klein & Murphy, 2001), where we see ambiguity effects for the ambiguous words themselves. This might be because participants are trying to integrate words both within and across trials in linguistic tasks, which would lead to task performance more similar to that observed for Elements 2 and 3 here (Klein & Murphy, 2001). Another possibility is that natural language tasks, as opposed to current artificial stimuli, engage in consistent, rapid, and automatic processing which results in detectable effects for the first element, whereas the slower and less natural processing of artificial stimuli do not elicit those effects.

We also investigated whether the slowdown for Element 3 was due to information integration per se, or was due to visual complexity. In a separate experiment not reported here due to space constraints, Experiment 2a was modified to have four symbols for all three sequence elements. We nevertheless still found significant slowdown between Elements 2 and 3 in polyseme and homonym sequences. This indicates that the slowdown observed in Experiment 2 was not solely attributable to differences in visual complexity for Element 3.

In contrast, the offline tests pointed to broadly similar performance regardless of the order in which the first two elements in the sequence were presented, with some detailed differences (e.g., changes in polyseme accuracy across Experiments 2a and 2b in sequence selection). This in turn suggests that the exact time-course of processing varies based on whether the more or less informative element is presented first, but the end result of processing is a relatively similar (although not identical) order-independent final representation.

Overall Comparisons

A striking difference between transitional probability and conjunctive probability sequences is in the long RT for Element 1. This may be explained by the ability of the first element to predict the following two elements in TP whereas both the first two elements need to be considered to predict the third in conjunctive probability.

In offline sequence completion, performance of high-entropy and low-entropy elements were similar for Experiments 2a and 2b in spite of their reversal in position within the sequence, supporting the hypothesis that performance on an element-level is tied to informativeness of the element. Generally, performances for offline tasks showed similar levels of accuracy across all experiments, suggesting that CPs do not pose much extra challenge in learning as compared to TPs.

General Discussion

Statistical learning is theorised to be a domain-general ability for detecting regularities across time and space, yet the bulk of extant research has focused on learning TPs between elements. This type of statistic, although clearly very useful for enabling some abilities like speech segmentation, is insufficient to understand other abilities, such as how words and contexts conjoin to evoke context-specific meanings in specific contexts (Swaab, Brown, & Hagoort, 2003). CPs, although certainly not capable of fully explain such behaviors, may be an alternative form of statistical computation that are critical for such information processing.

The present research merged a recent statistical learning paradigm, a self-paced learning task, with new statistical relationships among items that relate to CPs. Our results showed that CPs, like TPs, are learnable. By varying the amount of information content (entropy) in each position in the sequence, we were also able to ascertain that the order in which high- and low-entropy elements were presented in a sequence modulated online learning, but nevertheless resulted in similar patterns of performance in the offline test. Thus, the time-course of processing may differ based on the order in which information is presented (e.g., whether an ambiguous word like BAT precedes or follows a disambiguating context such as a discussion of SPORTS), but the end result of this processing is similar. Similarly, our manipulation of the relatedness between the "meaning" elements modulated performance in both the online and offline task, suggesting that the microstructure of each element can interact with the overall statistical regularities in the sequence. This suggests that multiple types of statistics among the individual elements of each sequence interact to determine overall performance.

This research represents an important proof of concept for how an alternative statistic than TPs can be learnt, and how such a structure could potentially interact with relatedness of interpretation to shape overall performance. In so doing, it opens up new possibilities for studying how simple statistical learning principles could interact with the rich structure of linguistic domains to explain at least some aspects of complex language behaviors such as context-sensitive meaning processing. As current models of statistical learning do not look at the problem of integrating constraints across elements, the current experiments can serve as a motivation to look at how this type of probability can be incorporated into such models. The ability to test even the domain-generality of some new language processes in a simple form is therefore very valuable. It also represents an important complement to existing methods using natural language, which have their own complexities in terms of controlling for confounding psycholinguistic properties. Having a new approach for developing convergent insights into statistical learning of CPs and other language abilities is therefore likely to be a powerful tool for advancing theory in related domains.

Acknowledgments

This work was funded by NSERC DG-502584 to BCA.

References

Armstrong, B. C., & Plaut, D. C. (2016). Disparate semantic ambiguity effects from semantic processing dynamics rather than qualitative task differences. *Language, Cognition and Neuroscience*.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological science*, *12*(6), 499–504.

Gómez, R. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*, 431-6.

Klein, D. E., & Murphy, G. L. (2001). The representation of polysemous words. *Journal of Memory and Language*, *45*(2), 259–282. doi: 10.1006/jmla.2001.2779

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*. doi: 10.1006/jmla.1996.0032

Siegelman, N., Bogaerts, L., Kronenfeld, O., & Frost, R. (2018). Redefining Learning in Statistical Learning: What Does an Online Measure Reveal About the Assimilation of Visual Regularities? *Cognitive Science*, *42*, 692–727.

Swaab, T., Brown, C., & Hagoort, P. (2003). Understanding words in sentence contexts: The time course of ambiguity resolution. *Brain and Language*, *86*(2), 326–343.