A generalization becomes suppressed over time in the context of exceptions

Karina Tachihara (tachihara@princeton.edu)
Department of Psychology, Princeton University, Princeton NJ 08544 USA

Kenneth A. Norman (knorman@princeton.edu)
Department of Psychology, Princeton University, Princeton NJ 08544 USA

Nicholas Turk-Browne (nicholas.turk-browne@yale.edu)
Department of Psychology, Yale University, New Haven, CT 06520 USA

Adele E. Goldberg (adele@princeton.edu)
Department of Psychology, Princeton University, Princeton NJ 08544 USA

Abstract

There has been a great deal of interest in how generalizations and exceptions are learned, with particular interest in how speakers learn to avoid overgeneralizations. Do overgeneralizations disappear only because exceptions become more strongly represented or does the generalization itself become suppressed? Novel labels were constructed by combining 56 syllables with one of two prefixes, and each label was assigned a unique image. Most labels with the first prefix were paired with images from a generalization category, whereas exceptional labels were paired with images from a different semantic category. All labels with the second prefix appeared with a third category ("baseline"). Participants used a computer mouse to choose one of two images for each label. Mouse-tracking results show that the generalization itself became suppressed over time in the context of exceptional labels. A post-test demonstrated that exceptions were learned with item-specific precision.

Keywords: language acquisition, generalization, exceptions, overgeneralization, mouse-tracking

Introduction

In order to speak a language fluently, it is critical to learn subclasses of exceptions within otherwise broad generalizations. For instance, in Spanish, words ending *–a* are generally grammatically feminine, but roughly half of the words that end in *–ma* are masculine (e.g., *el drama*). The present work investigates how these sorts of generalizations and exceptional subclasses interact with one another during the learning process. In particular, we investigate whether competition between a generalization and a subclass of exceptions persists to the same degree throughout learning.

Competition between generalizations and exceptions is widely recognized to affect language processing (Bates & MacWhinney 1987; Christiansen & Chater 1999; McClelland, & Rumelhart 1986; Goldberg 2019). However, less attention has been focused on how the process of learning exceptions might affect memory for the generalization. One possibility is that the generalization and exceptions are represented independently, and learning the exceptions has no effect on memory for the generalization. According to this perspective, the generalization and exceptions may operate in parallel and race to provide the correct form during production (Pinker 1999), or they may operate as sequential rules (Yang 2016). Both of these proposals are consistent with the idea that speakers learn to avoid overgeneralizations because exceptions become more strongly represented. No change in the representation of the generalization is required.

A third possibility we investigate here is that the generalization becomes suppressed in the context of exceptions. Support for this hypothesis comes from the literature on how competition between memories drives learning. Numerous studies have found that, when memories (semantic or episodic) compete, the "losing" memories (i.e., memories that are partially activated, but less than the memory that is fully retrieved) become harder to subsequently access, compared to memories that do not undergo competition. (Anderson et al., 1994; Anderson et al., 2000; Bäuml, 1998; Bäuml 2002; Johnson & Anderson, 2004; Levy et al., 2007; Murayama et al., 2014; Lewis-Peacock & Norman 2014; Kim et al., 2014).

For example, in Anderson et al. (1994), participants memorized a set of word pairs, some of which shared a semantic category (*fruit: orange; fruit: apple*) while other items were part of an unrelated category (*tool: hammer*). During the retrieval practice phase, participants were given a semantic cue and asked to recall a subset of the items (*fruit: ap___*). Note that the semantic category fruit can be expected to activate *orange*, but *orange* would lose in competition to *apple* because it is inconsistent with the partial cue "*ap___*".

That is, the cue ensures that *apple* wins in a competition with *orange* (and other prototypical fruits). At the final test phase, unsurprisingly, participants recalled practiced items (*apple*) best. Critically, items in the same category which were not themselves practiced (e.g., *fruit: orange*), had a lower recall rate than unrelated baseline items (*tool: hammer*), an effect known as retrieval-induced forgetting (RIF).

Anderson et al., (2000) emphasized the role of competition during retrieval in RIF. They found that simply repeating an item (e.g., *apple*) without the semantic cue (*fruit: ap____*) that could be expected to partially activate competitors such as *orange,* did not result in the subsequent suppression of *orange*. In this case where there was no competition-inducing cue, the repeated item (i.e., *apple)* was strengthened but the other word from the same category (i.e., *orange*) was not less likely to be recalled than words from other categories (like, *hammer*). These results demonstrate that it is not merely the strengthening of the more activated memory that resolves the competition. Rather, competition also leads to suppression of the less activated memory.

In the domain of language learning, we hypothesize that exceptions serve to delimit the domain of a generalization, suppressing its activation and carving out a space of their own so that the generalization and exceptions become more differentiated over the course of learning. The alternative hypothesis is that exception learning is the strengthening of the exception alone, with no change to the generalization. We aim to evaluate these hypotheses by exposing participants to a mini-artificial language that contained a generalization and a subclass of exceptions. We then used a mouse-tracking design, as it provides a sensitive way to detect competition between two alternatives in a forced choice task.

The mini-artificial language consisted of two prefixes and 56 syllables and images. One prefix appeared with 40 syllables paired with images of one semantic category (the generalization) and 8 other syllables paired with a second semantic category (the exceptions). The second prefix consistently appeared with 8 instances of a third semantic category and served as a baseline. For example, as presented in Figure 1, a subset of participants witnessed the prefix, *abber,* paired with 40 unique syllables and unique faces and 8 different syllables and unique scenes. The other prefix, *belling*, was then paired with 8 unique syllables and unique objects. The combination of semantic category (faces, scenes, objects) and prefix (*abber, belling*) was counterbalanced across participants, and additionally, the pairing of each syllable and image was randomized for each participant. However, for ease of description, we refer to the assignment of categories and prefixes represented in Figure 1 throughout the paper.

Participants were first exposed to all 56 <prefix+ syllable> pairs (hereafter, labels) and images. In the main task, participants heard each label and decided which of two images on the screen matched that label (vs. the other "lure" image) by using a computer mouse to move from the bottom of the screen to the chosen image (Spivey, Grosjean, & Knoblich, 2005; Spivey & Dale, 2006). These mouse-tracking trials were repeated over 8 blocks in order to investigate learning over time. Only correct trials are included in the main analysis. But the dependent measure used was deviation toward the distractor image (the "lure"), weighted by time, which captures the degree to which participants were lured by the *incorrect* category (Figure 2).

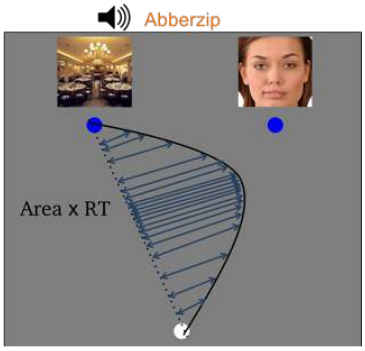

Figure 2: Example mouse-tracking trajectory sampled every 30 ms to determine the strength of the lure category (here, the face image).

Specifically, the distance between the cursor's position and a straight line to the correct response was measured at 30 millisecond intervals. To the extent that participants drew a relatively straight line from the start to the correct target, the deviation measure was low, indicating that the lure was not active in their minds. On the other hand, if participants drew an arc that trended toward the lure, we can conclude that the lure was activated by the label to some degree

Since our interest was in the relationship between a generalization and a subclass of exceptions, it might be tempting to focus on trials that included both an image from the generalization category and an image from the exception category. However, it would be impossible to determine in that case whether the trajectory was due to being lured by one image or by avoidance of the other. Specifically, an overgeneralization may be captured by a strong pull towards the generalization lure image or a lack of pull towards the correct exception image.

Therefore, in order to investigate how generalization activation changes without contamination from the lure of an exception image, a second trial type was introduced, "Scrambled-Image" trials (Figure 3). On these trials, participants were told to always select the scrambled image, regardless of what label was heard or which other image was available. For these

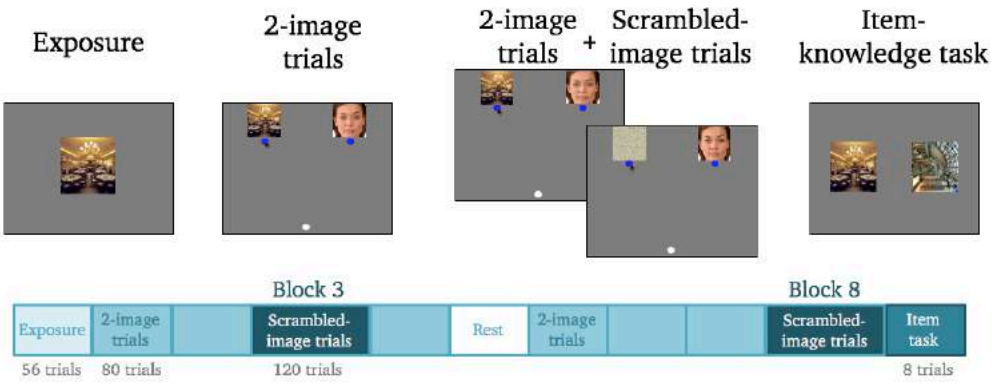

Figure 3: Exposure phase, followed by 8 blocks of 2AFC trials. Blocks 3 and 8 also contained scrambled-image trials in order to measure the strength of lures as directly as possible. Lastly, there was an item-knowledge task.

scrambled-image trials, the trajectory indicated how much participants activated the lure image category without the confound of preference toward the chosen image, since this was held constant across all scrambled-image trials. In order to avoid over-learning of the scrambled-image task (i.e., participants becoming good at going straight to the scrambled image without consideration of the lure), this type of trial was only included in blocks 3 and 8.

Blocks with two intact-image trials (blocks 1,2, & 4-7) each had 80 trials per block. Blocks 3 and 8 had two intact-image trials intermixed with scrambled-image trials for a total of 120 trials (Figure 3). Thus, in total, there were 720 trials, 11.11% of which were scrambled-image trials. Participants were not given any indication of the block structure of the task, except that they were given a rest after block 4, half way through the experiment.

By comparing performance on scrambled-image trials in blocks 3 and 8, the activation of lures over the course of learning can be detected. See Figure 3 for experimental design.

The scrambled-image trials, along with the mouse-tracking measure, allowed us to home in on the activation of a particular category for a particular label and how it changed over the course of learning. This enabled us to test the following hypothesis: a generalization becomes suppressed over time in the context of exception labels.

Method

The sample size and the main analysis were preregistered on Aspredicted.org, prior to data collection.

Participants

42 undergraduate students from Princeton University were compensated with course credit and up to an additional $5, depending on task performance.

Stimuli

The 2 prefixes (*abber* and *belling*) and 56 syllables (e.g., *zip, ber,* and *za)* were all phonotactically regular. The labels (prefix + syllable) were presented auditorily without pauses between the prefix and the syllable, and each lasted approximately 800 ms. Each scrambled image was created by scrambling the pixel locations of the lure image used in the same trial.

Procedure

Participants were given general instructions at the beginning as well as 6 practice trials for the 2AFC mouse-tracking task. They were told to pay attention to the pairing of the labels and images, but were not told about the structure of the stimuli (i.e., that the labels were a combination of a prefix and a syllable, nor the general distribution of categories). They were instructed to make their choices as quickly as possible and to move the cursor as directly to the target as possible while trying to avoid errors. The entire experiment lasted 1.5-2 hours, including a short rest period.

Initial exposure phase: Each label-image pair was presented once, for a total of 56 trials (40 generalization items + 8 exception items + 8 baseline items), with order of presentation randomized for each participant.

Mouse tracking task: Blocks 1-8: Participants were instructed to choose the image that matched the label they heard, except on scrambled-image trials in which they were instructed to always choose the scrambled image. For the intact-image trials, the two images always came from different categories, so participants could perform at ceiling by recognizing which category each label belonged to, without necessarily learning which face, scene or object each label corresponded to. For the scrambled-image trials, one of the images was created by scrambling the pixel

location of the other intact image. All other procedures were equal between intact image trials and scrambled-image trials.

The label was played through headphones. Once it was finished, participants could click the white button at the bottom of the screen, causing 2 images to be displayed. Participants then moved their cursor to the image that was associated with the label and clicked on the blue button underneath that image. In order to encourage participants to respond as quickly as possible, a score appeared on the center of the screen, calculated according to the trajectory of the mouse and speed of response. When the score was displayed, the incorrect image would disappear, leaving the correct image only. If participants had chosen incorrectly, they had to move their cursor to the correct image and click, before continuing to the next trial. After block 4, participants were given a mandatory 5-10-minute break before continuing with block 5.

Item-knowledge task: After the 8 blocks of the 2AFC task, participants performed a short task designed to test whether they had incidentally learned to associate particular exception labels with particular images within that category. The 2 images in this task were both instances of the exception category (e.g., 2 different scenes), requiring participants to identify the item-specific association of label and image. Participants were unaware they would be tested on item-specific knowledge for this task.

Results

All 42 participants exceeded the preregistered threshold of 75% accuracy on the mouse-tracking task ($M = 87.14$, $SD = 0.0085$), and none were excluded ($N = 42$). 3.25% of all trials were excluded because participants took > 2 seconds to click the start button or > 5 seconds to make a choice between images.

Accuracy on intact image trials

For trials in which participants decided which one of the two intact images matched the label they had heard, we can look at their accuracy against chance (50%) to see how well they knew the label-image pairings. Participants were above chance on all trial types in the first block after exposure ($t = 22.95$, $p < 0.0001$, $M = 0.89$), except for exception-label trials. On exception-label trials, participants heard an exception label and had to choose between an image from the exception category (e.g., scene, the correct choice) and the generalization category (e.g., face, the incorrect choice). Initially, accuracy on exception-label trials was significantly below chance (block 1), indicating that participants were systematically choosing the generalization image ($t = -2.13$, $p = 0.039$, $M = 0.43$). Accuracy for exception-label trials quickly rose, however, becoming significantly above chance in block 2 ($t = 2.43$, $p = 0.020$, $M = 0.59$). By block 8, accuracy for exception-label trials was as high as that for other trial types ($t = 1.30$, $p = 0.20$, $M = 0.93$ for exception trials, $M = 0.96$ for other trials).

Trajectory toward lure

The dependent measure for each trial was the area underneath the trajectory weighted by reaction time (area x RT). To calculate the area x RT, we compared the trajectory against the most direct, straight line connecting the starting point and the end point. The starting point was the position the participant had to click at the start of the trial and the end point was where the participant clicked when they made a choice (one of two blue circles). We measured points on the actual trajectory every 30 ms and calculated the distance between each of these points from the straight line. The sum of these distances is the area x RT. Note that the farther participants moved their cursor away from the straight line and the longer it stayed there, the higher the area x RT was. We had preregistered the dependent measure to be the maximum distance from the straight trajectory, and the result of the preregistered main analysis does not qualitatively differ when the maximum distance is used. However, after preregistering, we decided that area x RT was more appropriate and sensitive, allowing us to take both speed and deviation into account.

We report the results from the trajectory of the scrambled-image trials because it is the most direct measure of the activation of a category (i.e., the lure image category) given a label. For all analyses we used a maximal multilevel model with trial type or an interaction of trial type and block as the fixed effects and random intercepts and slopes for subjects and items where convergence would allow (Barr, Levy, Scheepers, & Tily 2013), using the lmerTest library (R Development Core Team 2008).

First, to confirm that highly activated lure images would indeed yield greater deviation and thus higher area x RT measures, we compared trials in which the label matched the lure image (e.g., the label was paired during the training phase with a specific scene and the lure image is that specific scene) and trials in which the label did not match the lure image or its category (e.g., the label was paired with a scene and the lure image is a face). As expected, we found that matched trials had a higher area x RT than unmatched trials ($\beta = -0.69$, $t = -11.06$, $p < 0.0001$).

Recall our hypothesis: that generalization activation (e.g., face image activation) would decrease over time for exception items (e.g., for labels that are paired with scenes). Thus, the critical preregistered comparison

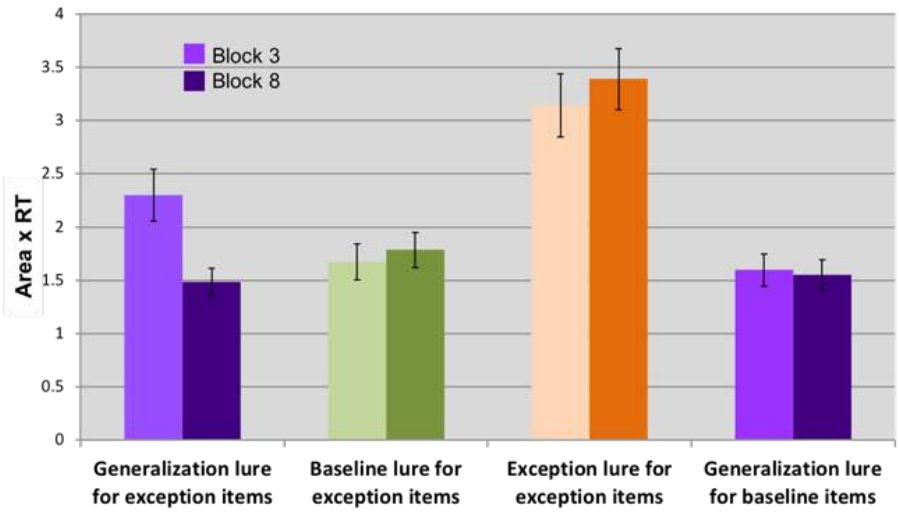

Figure 4: Deviation measure toward lure in block 3 (lighter color) and block 8 (darker shade), for exception-label trials with generalization image lures (left purple), exception-label trials with baseline lures (green), exception-label trials with exception lures (orange), and baseline-label trials with generalization lures (right purple). The correct choice in all cases was a scrambled image.

was the change of activation of the lure from block 3 to block 8 on trials when an exception-label was heard. We compared trials in which the lure image was the generalization (e.g., the label was paired with scene and the lure was a face image) against the baseline (e.g., the label was paired with a scene and the lure was an object image). The model found a significant interaction of trial type and block (β = 0.23, t = 2.66, p = 0.010). In other words, for exception items, the generalization activation became suppressed over time, more so than did baseline activation. In Figure 4, the far-most left panel shows the key generalization suppression from block 3 (light purple) to block 8 (dark purple). There is no suppression over blocks for baseline activation (green). Thus, generalization suppression cannot be attributed to general improvement over time or to general improvement on scrambled image trials.

Another critical part of the hypothesis is that the generalization was suppressed due to competition from learned exceptions. For exceptions to compete with the generalization, exceptions must be activated to some degree. In other words, exception-labels must be identified as exceptions and activate the correct exceptional category (scene) for competition to occur. Results additionally provide evidence that, as early as block 3, participants had learned which labels were exceptional. In particular, when an exception-label was heard, the matched exception image (scene) exerted a strong pull away from the scrambled image (third panel, light orange bar). In fact, the area x RT for matched exceptional images was higher than that for generalization images (face images) at block 3 (β = 0.42, t = 2.10, p =0.038). This means that the generalization suppression we observe occurred after participants had already learned the exceptions to some degree.

An alternative explanation for generalization suppression over time for exception-label trials could be that the generalization (e.g., face images) became less of a lure across the board, for exception items as well as baseline items. To investigate this possibility, we compared area x RT towards generalization lure images for exception-label trials against baseline-label trials. If (as hypothesized) generalization suppression is unique to exception labels (because of the competition from sharing a prefix), there should be no generalization suppression for baseline labels (where a prefix was never shared, and thus no competition took place). We again found a significant interaction of trial type and block (β = 0.19, t = 2.37, p = 0.018). In other words, generalization suppression over time was evident only in the context of exception items, not baseline items. In Figure 4, the far-most right panel shows no change of generalization activation over time for baseline items (purple bars). This also rules out the possibility that generalization suppression for exception items was specific to an image category (e.g., generally disliking faces over time).

Item-knowledge task

Despite high accuracy in the main task being achievable based purely on recognition that certain labels were exceptional (i.e., were associated with the non-dominant category for the prefix), the final task demonstrated that participants nevertheless learned with near-ceiling level accuracy which specific scene was paired with which specific label (M = .9494; t = 29.19, p < 0.0001).

Discussion and Conclusion

This experiment assessed how the activation of a generalization changed in the context of exceptions

over the course of learning. By using mouse-tracking to measure lure activation, we were able to isolate the activation of the generalization from the activation of the exception for a given label. Results demonstrate that the competing probabilistic generalization was a strong lure for the exceptions early on, but the generalization became suppressed over time in the context of exceptions. That is, the suppression of the generalization over time was evident only for exception labels. Because accuracy on exceptions was already high and exception lures were already even stronger lures early on, we suggest that the suppression was caused by competition from learned exceptions. Results of a post-test demonstrated that exceptions were learned with item-specific precision even though ceiling performance was possible by reliance on category membership only.

Importantly, our claim is not that learning eliminated all competition between generalizations and exceptions in this study. We know that comprehension is incremental, so we fully expect that listeners activated multiple options that were consistent with the input they witnessed until the point of disambiguation (Jurafsky, 1996; McQueen, 2007; Rayner & Clifton, 2009; Swinney, Prather, & Love, 2000); hearing *abber* should trigger a competition between exceptional items (e.g., *abber zip*) and other items that begin with the same prefix (e.g., *abber fep*), even after learning takes place. Rather, our main hypothesis pertains to what happens after the disambiguating syllable (*zip*) is heard: Would learning of exceptions affect the activation of the generalization, specifically in the case of exceptions like *abber zip*? We found that it did: the generalization became a less powerful lure as exceptions became more easily identified.

This work was motivated, in part, by the effects of competition on memory observed during studies of *retrieval-induced forgetting* (RIF). Consistent with RIF findings in the memory literature, the linguistic generalization became suppressed ("forgotten") in the context of exceptions. At the same time, it is important to point out a key difference between our study and the way RIF is usually tested. Most RIF studies look at final recall to measure memory performance. Our study, on the other hand, considered the change in activation over the course of learning. This difference led us to use a different baseline for determining whether suppression occurred. In standard RIF studies, suppression is measured by how much lower the memory for competing items is, compared to baseline items which had not been in competition with the practiced items. In our studies, suppression was measured by how much lower the activation for the generalization became over time. We found that generalization activation significantly decreased over time for exception items, much more than it did for baseline items. However, we did not find that activation levels of the generalization fell below baseline activation; as such, we did not find RIF in the classic sense. Nonetheless, our results are consistent with the idea that competition leads to suppression of the less activated memory.

We selected the "prefix plus syllable" structure for the labels to allow for prediction of the category given the prefix. The prefix plays the role of a classifier (i.e., a grammatical element that selects for nouns of certain semantic categories; Dixon 1986) As noted in the introduction, linguistic categories, including classifier categories, often have subclasses of exceptions, as in the present experiment. However, the finding in this study is not, in principle, specific to words. For example, similar competitive mechanisms may serve to suppress grammatical overgeneralizations through what has been called *statistical preemption* (e.g., Goldberg, 2019; Perek & Goldberg 2017; Robenalt & Goldberg, 2015). Future work will build on these results to explore how generalizations and exceptions compete in other domains, how the underlying neural representations change, and how these competition-driven changes relate to behavior.

References

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting - Retrieval Dynamics in long-term-memory. *Journal of Experimental Psychology-Learning Memory and Cognition*, *20*(5), 1063–1087.

Anderson, M. C., Bjork, E. L., & Bjork, R. A. (2000). Retrieval-induced forgetting: Evidence for a recall-specific mechanism. *Psychonomic Bulletin and Review*, *7*(3), 522–530.

Bates, E., MacWhinney, B., & MacWhinney, B. (1987). Competition, variation, and language learning. *Mechanisms of language acquisition*, 157-193.

Bäuml, K. (1998). Strong items get suppressed, weak items do not. *Psychonomic Bulletin and Review*, *5*(3), 459–463.

Bäuml, K.-H. (2002). Semantic generation can cause episodic forgetting. *Psychological Science*, *13*(4), 356–360.

Christiansen, M. H., & Chater, N. (1999). Connectionist natural language processing: The state of the art. *Cognitive science*, *23*(4), 417-437.

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. M. (2001). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes*, *16*(5-6), 507-534.

Dixon, R. M. (1986). Noun classes and noun classification in typological perspective. *Noun classes and categorization*, 105-112.

Goldberg, A.E. (2019). *Explain me this: creativity, competition and the partial productivity of constructions.* Princeton: Princeton University Press.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & psychophysics*, *28*(4), 267-283.

Johnson, S. K., & Anderson, M. C. (2004). The role of inhibitory control in forgetting semantic knowledge. *Psychological Science*, *15*(7), 448–453.

Jurafsky, D. (1996). A probabilistic model of lexical and Syntactic Access and Disambiguation. *Cog. Sci. 20*(2), 137–194.

Kim, G., Lewis-Peacock, J. A., Norman, K. A., & Turk-Browne, N. B. (2014). Pruning of memories by context-based prediction error. *Proceedings of the National Academy of Sciences*, *111*(24), 8997–9002.

Levy, B. J., McVeigh, N. D., Marful, A., & Anderson, M. C. (2007). Inhibiting your native language: The role of retrieval-induced forgetting during second-language Acquisition. *Psychological Science*, *18*(1), 29–34.

Lewis-Peacock, J. A., & Norman, K. A. (2014). Competition between items in working memory leads to forgetting. *Nature Communications*, *5*(5768), 1–10.

McClelland, J. L., Rumelhart, D. E., & PDP Research Group. (1986). Parallel distributed processing. *Explorations in the Microstructure of Cognition*, *2*, 216-271.

McQueen, J. M. (2007). Eight questions about spoken-word recognition. *The Oxford Handbook of Psycholinguistics*, 37–53.

Murayama, K., Miyatsu, T., Buchli, D. R., & Storm, B. C. (2014). Forgetting as a consequence of retrieval : A meta-analytic review of retrieval-induced forgetting. *Psychological Bulletin*, *140*(August), 1383–1409.

Perek, F., & Goldberg, A. E. (2017). Linguistic generalization on the basis of function and constraints on the basis of statistical preemption. *Cognition*, *168*, 276-293.

Pinker, S. (1999). *Words and rules: The ingredients of language*. Basic Books.

Rayner, K., & Clifton, C. (2009). Language processing in reading and speech perception is fast and incremental. *Bio. Psych.*, *80*(1), 4–9.

Robenalt, C., & Goldberg, A. E. (2015). Judgment evidence for statistical preemption: It is relatively better to vanish than to disappear a rabbit, but a lifeguard can equally well backstroke or swim children to shore. *Cognitive Linguistics*, *26*(3), 467-503.

Spivey, M. J., & Dale, R. (2006). Continuous dynamics in real-time cognition. *Current Directions in Psych. Sci.*, *15*(5), 207–211.

Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *PNAS*, *102*(29), 10393–10398.

Swinney, D., Prather, P., & Love, T. (2000). The time course of Lexical Access and the role of context. In *Language and the Brain: Representation and processing* (pp. 273–292). New York: Academic Press.

Yang, C. (2016). *The price of linguistic productivity: How children learn to break the rules of language*. MIT Press.