Linking Memory Activation and Word Adoption in Social Language Use via Rational Analysis

Jeremy R. Cole, Moojan Ghafurian, and David Reitter
jrcole,moojan,reitter@psu.edu
The Pennsylvania State University
University Park, PA USA

Abstract

This paper investigates how cognition facilitates the adoption of new words through a study of the large-scale Reddit corpus, which contains written, threaded conversations conducted over the internet. Parameters for the cognitive architecture are estimated. Using ACT-R's account of declarative memory, the activation of memory chunks representing words is traced and compared to usage statistics sampled from a year of data. Potential values for decay and retrieval threshold are identified according to model fit and growth rates of word adoption. The resulting estimate for the decay parameter, d, is 0.22, and the estimate for the retrieval threshold parameter, rt, lies between 3.4 and 4.5.

Keywords: neologisms, retrieval threshold, decay

Introduction

Language is a communication system that varies among speakers and is constantly changing. Naturally, language occurs in the context of social interaction, and large-scale datasets reflecting language use are a good opportunity to study individual cognition in the social context. It is this context that the cognitive architecture may have evolved to serve.

The aspects of the architecture most linked to the adoption of new words among individual language users is are declarative memory formation retrieval. English is a productive language: new words are invented frequently. In fact, the rate of new word formation has increased in the past century (Lehrer, 2006). Newly introduced words might be used for only a short period of time or may last longer and contribute to large-scale language change. This process relies on speakers taking liberties with their word choice and on speaker communities that facilitate and accept the use of novel words.

In this paper, we model word choice and exposition to words as the result of declarative memory activation (Anderson and Schooler, 1991). This lets us study the cognitive architecture in the context of the social environment, as it presents itself in a very large corpus of web-forum dialogue. As a result, we are able to derive rational parameters for the ACT-R declarative memory module.

Lexical change has been studied experimentally. For example, *naming games* have proven to be a fruitful way to elicit change (e.g., Baronchelli 2011). The dispersion of new ideas has also been observed in large-scale data as well. Hashtags in Twitter are a good example of neologisms that represent memes. Their dispersion dynamics can be surprising in that they appear to be different depending on the topic (Romero et al., 2011). For controversial topics, e.g. in politics, repeated exposure keeps achieving additional adoption (*complex contagion*).

However, to our knowledge, little work has studied word adoption at an individual level through cognitive modeling. We take this as an opportunity to employ rational analysis to fit architectural parameters. While Anderson and Schooler (1991) touched on this, determining certain features of memory that must be true in order to process newspaper headlines. Relatedly, we model the state of memory directly to determine the optimal fit of parameters based on the data.

For a word to be used spontaneously, it must have high enough *activation* to be retrieved. This presents a bit of a conundrum, and perhaps an explanation for why this level of analysis has been avoided: to more highly activate a word, it must be presented, but for it to be presented, someone must successfully retrieve it. Nonetheless, one can assume that there are some people, the originators, for whom the word is more highly active. Then, as these people are relatively few in number, we can still measure the approximate activation for the adopters. This allows us to find the threshold for adoption, and thus guess at the threshold for retrieval.

In this paper, we thus present a simple cognitive model of word adoption. It uses a computational measure of activation and a corpus of the *Reddit* web forum to investigate the role of memory in word adoption. Beyond word adoption, we are interested in using measures of activation to compare to more empirical results, such as frequency. By using such empirical measures across a wide dataset, we can measure accurate values for certain parameters of ACT-R that have only been guessed at based on small-scale experimental results (Anderson, 1983). In particular, we focus on fitting the value of d, the decay parameter, and estimating the value of rt, the retrieval threshold parameter.

There are a few related topics that converge to our research questions. In particular, we are interested in the cognitive mechanisms that cause the adoption of new words (or neologisms) or new ideas in general, as well as the ability to use big data to provide evidence toward parameters in cognitive models. Lastly, most models ultimately provide information about declarative memory elements that already have been presented. This model's

novelty, in part, is due to its evaluation of new elements and an evaluation on corpus data.

Related Work

Beyond naming games, which have focused primarily on social factors, there are a few studies on the impact of cognitive factors on word adoption. For instance, Gilhooly (1984) showed that age of acquisition is more important than 'residence times' in naming times. They likewise relied on new words based on their introduction to language. Indeed, age of acquisition has been related to several such experimental paradigms and in many other studies (e.g., Morrison and Ellis 1995). While these studies are interesting, they have not focused on how such factors impact the adoption of new words, just how well they nestle in a single person. Other studies we know of that take memory into account at all also do not take adoption into account (e.g., De Vaan et al. 2007).

Most work that is focused on word adoption at a large scale has focused on *lexical innovation*, which normally has a focus on word forms, rather than memory and time course (e.g., Baayen and Renouf 1996). However, an important component of word adoption is not just whether the word form is easy to learn, but whether it can be retrieved from memory at all.

Previous work has focused on the relationship between memory and traditional measures of activation found in corpora, such as recency and frequency (e.g., Anderson and Schooler 1991). While that work was fundamental, it did not develop estimates for modern ACT-R parameters.

This calls for a cognitive model, as some value of activation should correspond to the retrieval threshold. While this value is used in ACT-R, to our knowledge, there are no papers estimating its empirical value in any field, and we are certainly aware of none estimating it in language.

Cognitive models of language are of course not new. Both comprehension (e.g., Lewis and Vasishth 2005; Ball et al. 2010) and production have been explored (e.g., Guhe 2009; Reitter et al. 2011). Language acquisition has also been explored (e.g., Dörnyei 2009), though it has mostly focused on second language acquisition. This is because it is difficult to acquire realistic human language data at acquisition time. Cognitive models of language acquisition without such strict constraints are much more common (e.g., Pinker and Prince 1988). By focusing on new words, we provide a possible workaround. By using a corpus, we have a lot of data in order to look at certain effects.

Methods

In general, our evaluation relies on comparing the data created by our model of activation with the human data from the corpus. This type of evaluation lets us fit against a large amount of data, not only confirming previous findings about ACT-R but tuning and estimating certain parameters.

Data: Reddit Corpus

Our data set consists of approximately 426GB of Reddit data, ranging from the year 2012 to the year 2014. Reddit.com is a community-driven news aggregation website that mostly contains discussions and ratings on a variety of topics (Bergstrom, 2011). The various communities the topics are organized around are called *subreddits*.

After the submission, people can reply with their thoughts in a *comment*. Users can also comment on these comments. We study these comments. Before applying any of our analysis, we filter out comments in subreddits with a small number of users (defined as 500). As anyone can make a subreddit and invite their friends to join, we wanted to avoid small subreddits that may more closely resemble social networks than communities.

What constitutes a new word?

As discussed, our data spans 2012-2014. In this sense, we came up with a simple way to determine if a word is new: it did not occur in 2012, but it did occur in 2013 or 2014. To ensure we excluded non-linguistic or pseudo-linguistic elements (such as hyperlinks), we excluded every token that did not entirely consist of alphabetic characters. To ensure we excluded typos or words that only had meaning in a single conversation, we used a simple arbitrary cutoff of one hundred occurrences. We claim that these three requirements are sufficient to define a new word, or a neologism. Some example words can be found in Table 1. In total, we found 3545 words matching these criteria.

There are two important limitations to this. Some elements of this set of words only have meaning to members of that subculture; some of them may have even fallen out of use already. Secondly, some of these words originate from a culture external to Reddit. In some of these cases, the usage of the words is still novel: Square Cash, a financial product, was frequently referred to as *squarecash* by Reddit users. Others, however, are strictly adoptions, such as Chromecast. Thus, we will refer to these as *first adoption* events. However, the cutoff for number of occurrences does indicate that these are true adoptions, not simply one-off usages.

While some of the first adoption events are origination events, all of them are a discussion of something new. The first discussion of a new idea has social consequence. In Reddit, people receive both explicit and implicit rewards for social acceptance, through the karma mechanism. We will use *adoption* to refer to any usage of a new word by a subreddit, using origination or first adoption for the first subreddit to adopt it, and *later adoption* for later usages.

Table 1: A table containing examples of new words, along with the subreddit it first appeared in, and a subreddit that it appeared in later. Both of these events are treated the same in our model.

Word	First Adopter	Later Adopter
dogetips	dogecoin	funny
misanderkirby	AdviceAnimals	AskReddit
peshka	gaming	Warthunder
gamecribs	leagueoflegends	counterstrike
squarecash	economy	Bitcoin
watchapps	pebble	Android

Cognitive Model

We see the adoption process as one that is governed by declarative memory (DM). A word is added to the lexicon (in DM), and through repeated presentations, it becomes available. We conjecture that initial use of the word is aided by short-term memory, from direct copying, or aided by cues (which spread activation, if seen from an ACT-R perspective). At some point, activation of the memory trace in the modeled individuals reaches the point where this word is retrievable without the help of cues. This *retrieval threshold*, as well as the function governing the gradual rise in activation, are central to this model, and we will estimate their parameters from the data.

We compute the activation of adopted words using the base-level learning equation defined originally by Anderson (1983).

$$bll(x) = \log\left(\sum_{i \in P_x} t_i^{-d}\right)$$

In this equation, x represents any symbol, a word in our case, and P_x refers to the list of x's presentations. So t_i is the time from that presentation to the present. Naturally, for something with as many presentations as any given word, it is infeasible to computationally manage that sum. However, the full equation can be approximated using only the total number of presentations and the k most recent presentations and $n_x = |P_x|$ (Petrov, 2006).

$$bll(x) \approx \log\left[\sum_{i}^{k} t_i^{-d} + \frac{(n_x - k)\left(t_{n_x}^{1-d} - t_k^{1-d}\right)}{(1-d)\left(t_{n_x} - t_k\right)}\right]$$

Petrov (2006) shows that the equation is close even for $k = 1$. As the amount of events in the Reddit corpus is very large, computing the many previous events is computationally expensive. Thus, we relied on this approximation and only kept track of the single most previous event. Note that causes the left sum to collapse to t_k^{-d}.

Table 2 gives brief descriptions of how each parameter was computed. For the constant d, we initially examine two values: 0.5, the ACT-R default (Bothell, 2004), and 0.16, as found by Vasishth and Lewis (2004).

Activation, in ACT-R, is composed of the base-level learning function (as above), in addition to spreading activation from cues and noise.

Table 2: The parameters of our activation equation and a description of how we computed them

Parameter	Description
n_x	The total number of occurrences of that word across Reddit
t_k	The time in between the current usage and the previous usage
d	The decay parameter, 0.5 or 0.16
t_{n_x}	The amount of time since the first usage of the word
k	1

Retrieval Threshold

The retrieval threshold rt defines the point of total activation for a memory trace to be retrievable. Obviously, many assumptions influence this parameter, to include how many times we assume the item to have been used in the past, outside of the context of the experiment at hand. As a consequence, no canonical value for this parameter is available. However, by looking at new words, which not based on past experience, and influenced less by external influence, we may be able to approximate this threshold.

Filtering the data

In order to get a realistic estimate of rt, we had to look at the pattern of the data. In particular, we wanted to see at what point words were adopted. However, the time course data taken naively is somewhat biased: because each word is adopted at a different point, and our data is only for just over 400 days, the number of words being evaluated is different at each day. Thus, to get the full range of effects while still avoiding bias, we only included words with over 400 days of data, and excluded all data beyond 400 days.

Relating adoption to declarative retrieval

Defining exactly what it means when the word is 'adopted', and thus has activation above rt is non-trivial, which is likely why there is so little information on it throughout the literature. However, our hypothesis was fairly simple: once activation is high enough that retrieval is possible, the frequency of usage will rapidly expand, as it usage no longer relies on referencing external sources. We will estimate that point by observing the pattern of results and finding where the derivative increases.

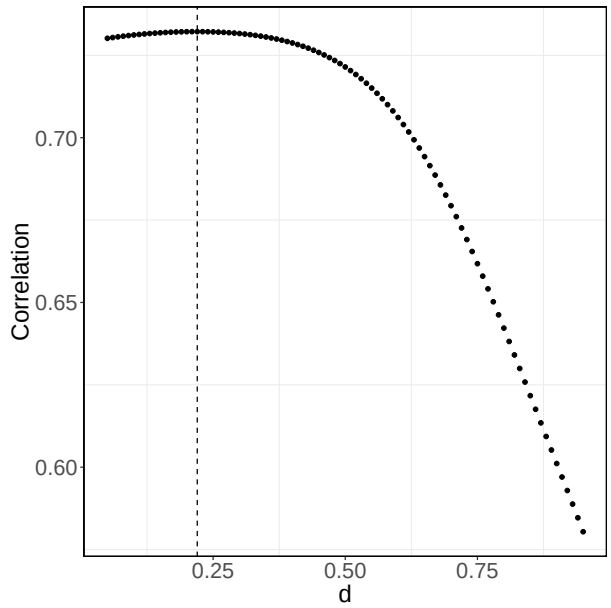

Figure 1: Correlations between word usage and the calculated activation for different d values, ranging from 0.05 to 0.95.

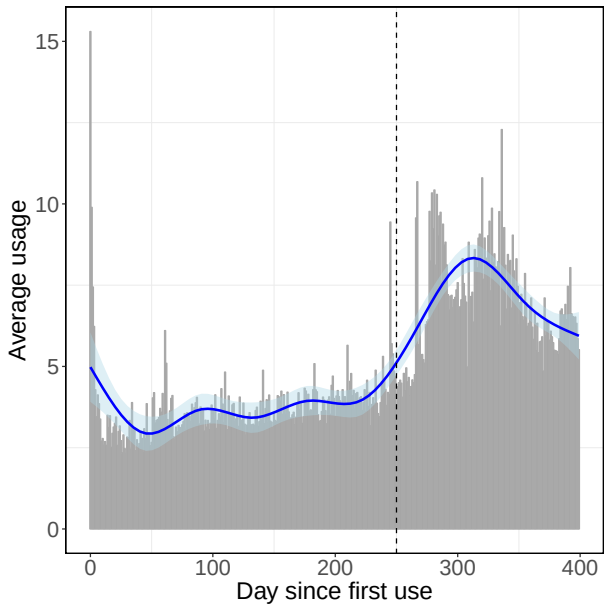

Figure 2: Average usage of new words per day, over time. Day 0 represents the day on which the word was first adopted. The dotted line marks the day where the derivative has clearly changed, around day 250. This inflection point represents the adoption event.

Fitting the decay parameter

In order to fit the decay parameter, we show which measure of activation, computed as described earlier, best predicts usages per day. The usage of each word every day is an empirical metric that should show how active that word actually is. We ask whether the ACT-R default (.5) or the value found by Vasishth and Lewis (2004) (0.16) yields a better fit, or if a different value would be found altogether. A grid search between between .05 and .95 was used, optimizing the activation's correlation with usages per day while using that value.

Results

By closely examining the data, we are able to see a clear inflection point for rt, as well as a pattern in the fit of activation.

Decay parameter

The activation for each word occurrence was calculated for different values for d. Figure Figure 1 shows the correlation between activation and observed word usage as a function of d. The correlation peaks at 0.22 (see Figure 1). Note that this methodology is approximate and assumes, e.g., $k = 1$. So, this disagrees with a value of 0.5, but, largely, agrees with 0.16 reported in the literature.

Retrieval Threshold

After showing the pattern of usages per day over time, there is a point where the function changes from oscillating but linear to a more exponential curve. In other words, we see a change in the derivative as the word is 'adopted', leading to larger gains as the word is able to be used more freely. This is around 250 days in, as shown in Figure 2. Then, looking at the activation over time

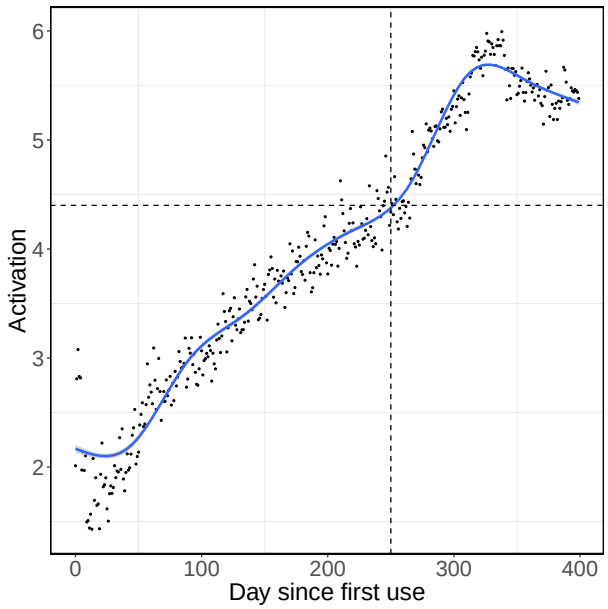

Figure 3: Base-level activation for $d = 0.16$ calculated for each word occurrence over time. Day 0 represents the day on which the word was first adopted. The vertical dotted line is at the same day as the inflection point shown in Figure 2; the horizontal line shows the activation for this value of decay, about 4.4. This value represents a possible value for rt.

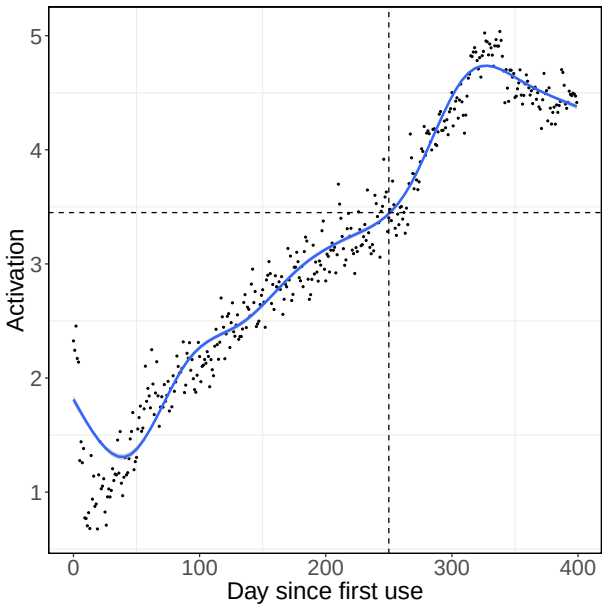

Figure 4: As Figure 2, for $d = 0.22$. The inflection is near activation 3.45. This value represents a possible value for rt.

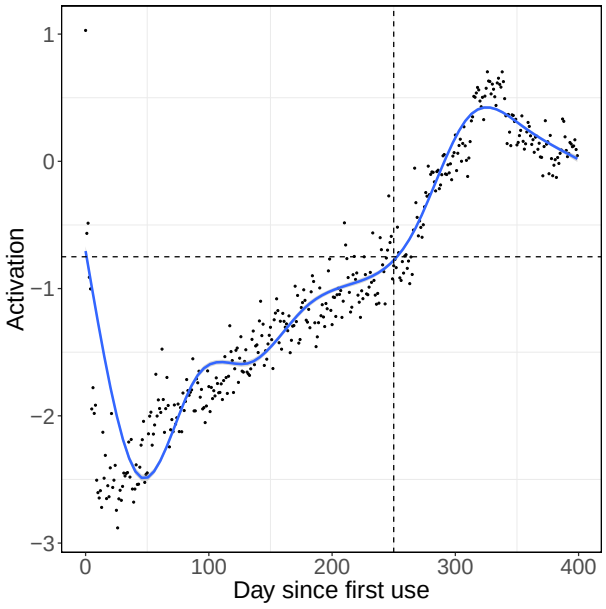

Figure 5: As Figure 2, for $d = 0.50$. The vertical shows the time of the same inflection point shown, here at activation -0.75

we are able to find an approximate value for activation where that changes, approximately 4.4 (see Figure 3). However, this is based on $d = 0.16$, we also estimate it at about 3.45 for $d = 0.22$, our own empirically found value (see Figure 4). Lastly, we compute it for the ACT-R default of $d = 0.5$, and find it to be -0.75 (see Figure 5). In general, the values found for activation for $d = 0.5$ also suggest it is not particularly suitable, as a negative activation should not be retrievable at all. These values correspond to reasonable guesses for rt. In particular, based on our methodology, we do not claim either 250 or the values for rt are the exactly correct values; however, based on the results of this study, they present reasonable constraints for an estimation of rt. In particular, we chose 250 as a point that is clearly starting an ascent and is significantly above the earlier trend.

Discussion

This paper has used ACT-R memory retrieval on data that reflects long-term language use in a social context. With this, we examine two critical parameters in declarative memory retrieval: decay (d) and the retrieval threshold (rt).

With this idea, we follow the idea of rational analysis: can we observe environmental data to draw conclusions about the individual cognitive system, assuming that it has evolved to be optimally adapted to process information from this environment while contributing to the production of such data in the first place. However, what is perhaps more unique to our approach is that we observe language behavior in a large-scale and long-term social context.

As for d, we obtain a best fit at a very different rate of decay than what is observed in controlled behavioral experiments. Of course, many standard experiments on memory retrieval use words, so language is not necessarily unique to the data in the present study. For language in context, as in the Reddit data, the slower decay of language could be due to the heavy semantic relatedness in language, which causes constant spreading activation. Even new words are largely derivative of old ones, borrowing phonetics, ideas, roots, or at least lexicography. Naturally, when dealing with models over time courses that make sense for language, the difference between 0.50, and 0.22 is substantial. We provide additional evidence that the value could be different.

One explanation of the slower decay that we observe may be reinforcement through cognitive function that is not observed: in other words, people do not write a word in a Reddit post every time they think of it. The other consideration is the time-course of word adoption: we have examined language use through about one year (the *social band*, Newell, 1990), while ACT-R's declarative memory framework is currently best suited to seconds and minutes (the *cognitive band*).

Still, an important constraint is that Reddit consists of written language and conversations can span several days. This could be a possible problem, as it is unclear how applicable forum discourse is to laboratory studies. A similar study performed on a corpus of real-time communication could be informative.

Our method for fitting rt opens up many interdisciplinary questions beyond the scope of this exploratory study. What is the range of the inflection point when

considering multiple samples from the corpus, and from other corpora? Are there meaningful *bands* (to use Newell's term) identifiable in the behavior of activation before day 50 and after day 300? Does language use in context not follow the patterns of cue-based memory retrievals found in dedicated experiments? How does the socio-informational network contribute to a changing inflection point and a critical mass necessary for contagion? Nonetheless, we acknowledge the limitations in the approach for measuring rt, though leave it to future work to determine a more precise measurement, perhaps based on fitting a model of retrieval to the corpus data.

Conclusion

In this paper, we use a cognitive model of memory that models the process of learning new words. By evaluating the model on corpus of social, contextual language use, we are able to model large amounts of human data, which gives us insight into the process and lets us examine the ACT-R model of memory itself. By comparing against an empirical measure of 'activation', we are able to correlate activation was computed by ACT-R in order to determine a reasonable value for d in language. We are also able to compute a reasonable estimate for rt, a parameter that has yet to be fitted in language or other domains. Specifically, we found that the bounds for a retrieval threshold we found lie somewhere between 3.4-4.5; the decay d at .22 or lower – unlike in many other studies.

Acknowledgements

This project was supported by the National Science Foundation (BCS-1457992 and IIS-1459300).

References

Anderson, J. R. (1983). A Spreading Activation Theory of Memory. *Journal of Verbal Learning and Verbal Behavior*, *22*, 261–295.

Anderson, J. R. & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, *2*(6), 396–408.

Baayen, R. H. & Renouf, A. (1996). Chronicling the times: productive lexical innovations in an english newspaper. *Language*, 69–96.

Ball, J., Freiman, M., Rodgers, S., & Myers, C. (2010). Toward a functional model of human language processing. In *Proceedings of the 32nd annual meeting of the Cognitive Science Society* (pp. 1583–1588). Portland, Oregon.

Baronchelli, A. (2011). Role of feedback and broadcasting in the naming game. *Physical Review E*, *83*(4), 46–103.

Bergstrom, K. (2011). Don't feed the troll: shutting down debate about community expectations on Reddit.com. *First Monday*, *16*(8).

Bothell, D. (2004). *ACT-R 6.1 reference manual*. Tech. Rep., 2004.

De Vaan, L., Schreuder, R., & Baayen, R. H. (2007). Regular morphologically complex neologisms leave detectable traces in the mental lexicon. *The Mental Lexicon*, *2*(1), 1–23.

Dörnyei, Z. (2009). *The psychology of second language acquisition*. Oxford: Oxford University Press.

Gilhooly, K. (1984). Word age-of-acquisition and residence time in lexical memory as factors in word naming. *Current Psychology*, *3*(2), 24–31.

Guhe, M. (2009). Generating referring expressions with a cognitive model. In *Proceedings of the Workshop Production of Referring Expressions: Bridging the Gap between Computational and Empirical Approaches to Reference*. Amsterdam, Netherlands.

Lehrer, A. (2006). Neologisms. In K. Brown (Ed.), *Encyclopedia of language & linguistics (second edition)* (Second Edition, pp. 590–593). Oxford: Elsevier.

Lewis, R. L. & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*(3), 375–419.

Morrison, C. M. & Ellis, A. W. (1995). Roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(1), 116.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Petrov, A. A. (2006). Computationally efficient approximation of the base-level learning equation in ACT-R. In *Proceedings of the seventh international conference on cognitive modeling* (pp. 391–392).

Pinker, S. & Prince, A. (1988). On language and connectionism: analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1), 73–193.

Reitter, D., Keller, F., & Moore, J. D. (2011). A Computational Cognitive Model of Syntactic Priming. *Cognitive Science*, *35*(4), 587–637.

Romero, D. M., Meeder, B., & Kleinberg, J. (2011). Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on Twitter. In *Proceedings of the 20th international conference on World Wide Web* (pp. 695–704). ACM. Hyderabad, India.

Vasishth, S. & Lewis, R. L. (2004, July). Modeling sentence processing in ACT-R. In *Proceedings of the ACL workshop incremental parsing: bringing engineering and cognition together* (pp. 82–87). Barcelona, Spain.