The Timing of Lexical Memory Retrievals in Language Production

Jeremy R. Cole and David Reitter
College of Information Sciences and Technology
The Pennsylvania State University
University Park, PA
jrcole,reitter@psu.edu

Abstract

This paper explores the time course of lexical memory retrieval by modeling fluent language production. The duration of retrievals is predicted using the ACT-R cognitive architecture. In a large-scale observational study of a spoken corpus, we find that language production at a time point preceding a word is sped up or slowed down depending on activation of that word. This computational analysis has consequences for the theoretical model of language production. The results point to interference between lexical and phonological stages as well as a quantifiable buffer for lexical information that opens up the possibility of non-sequential retrievals.

1 Introduction

Speech varies greatly in fluency, and some of its speed variation can be traced to the utterance spoken (Jespersen, 1992). Low-frequency words, for instance, are known to slow down speech (e.g., Bell et al., 2009). Variables correlated with fluency give valuable cues to the architecture of the language processing system. However, a model to explain these data has yet to emerge.

In this paper, we propose a cognitive model of fluency, in which lexical memory retrievals may explain some of the variability in speech rates. In particular, frequency, context and recent uses together have the potential to quantify retrieval delays through *activation* (Anderson, 1991). Activation, in its most common usage, refers to the way nodes in semantic networks become easier to retrieve after adjacent nodes have been activated, typically through a presentation (Collins and Loftus, 1975). In particular, activation makes a direct claim that more highly activated words require less time to retrieve, and vice versa (Anderson, 1983).

The language production process as a whole likely requires some amount of sequential processing. For instance, the standard model proposes that an idea is generated, lexicalized, grammatically and morphologically encoded, and only then phonologically encoded (Bock and Levelt, 2002). Still, most models of language production presuppose some amount of planning of output (e.g., Pickering and Garrod, 2013), so we could instead divide language production into planning this output and the actual process of outputting. The overlap and relationship of these processes is not fully understood, but given that most output is likely planned, the scale at which the planning takes place and the amount of time between planned output and the actual process of outputting remains unclear. However, if interactions between processes are observed, then we can likewise see when they overlap in time.

To summarize, we are suggesting that some of the variance in speech rate is not due to the linguistic properties of the words currently or about to be outputted, but the words still in the planning phase. We propose a model that uses a buffer of several words between initial retrieval and output, during which grammatical and morphological encoding take place. We examine this by calculating retrieval activation for a word and evaluating the influence of that activation on the empirical speech timing several words beforehand, using the Switchboard corpus. The effect of activation is distributed over preceding words in a way that is characteristic of a shared-resource, buffer-based account of language production.

2 Related Work

2.1 Stages of Language Production

Grammatical encoding can be divided into *functional* and *positional* processing steps (Bock and Levelt, 2002). The functional step selects lexical items and assigns functions, while

Proceedings of NAACL-HLT 2018, pages 2017–2027
New Orleans, Louisiana, June 1 - 6, 2018.

the positional step then combines the items to produces constituents. In our account, we expect that these mutually dependent steps work in parallel.

An important early part of functional processing retrieves lexical information, which we will examine in this paper. We evaluate the consequences of lexical access, which is assumed to be affected by the cost associated with any retrieval from declarative memory. Much discussion in this area has concerned the question whether lexical access happens in a single stage (Dell et al., 1997) or in multiple stages and overlaps with grammatical encoding (Caramazza, 1997; Roelofs et al., 1998; Caramazza, 2006). Here, we follow ACT-R's serial and partially symbolic nature, which in turn leads to some theoretical commitments to non-parallel processing: language production is staged and discrete. Nonetheless, each stage can be composed of several steps, and steps from syntactic and phonological processing likely interleave. This is compatible with empirical findings and the overall theoretical debate (Ferreira and Slevc, 2007). The precise timeline of processing is unclear, but as we will argue in this paper, large-scale speech data can give us usable clues to that effect.

2.2 Incrementality in Language Production

The second issue we address concerns the timing of memory retrievals, which is also related to the idea of incremental processing. It is a commonly implied assumption that language processing proceeds incrementally. In grammatical encoding, this property concerns when and in which order syntactic choices are made. For instance, all of them could be made before phonological processing starts (non-incremental case), or they could be made in order as necessary. Existing high-level models of language production proceed incrementally at various steps in a chain of content selection, aggregation and sentence realization (e.g., Bock and Levelt, 2002; Guhe, 2007).

Ferreira (1996) makes an argument for incrementality, based on the observation that competitive syntactic alternatives facilitate production rather than making it more difficult. An incremental account of sentence realization would predict such an effect, as syntactic "flexibility" introduced by the alternatives makes it easier to find a workable syntactic decision. By contrast, without incremental commitment to each structure, competing material slows down the process, because it would lead to combinatory explosion. However, later results establish nuance. Ferreira and Swets (2002) show that incremental production is possible, but it is "under strategic control"; it depends on semantic information, and it could be modulated by external factors, such as stress.

If processing were fully incremental, then it would follow that lexical memory retrievals are also fully incremental. The order words are retrieved in would be the same as the order words are eventually outputted in. However, if other features modulate this, then it would imply that incremental processing is instead variable, as suggested by earlier accounts.

2.3 Speech Rates

Several studies have illustrated the effects of frequency, recency, and context (Bell et al., 2009; Arnon and Priva, 2014) on speech rates. These studies motivated our modelling choices, as recency, frequency, and context are also the key components of the ACT-R theory of memory.

Recent research has found a correlation between rate of speech and the information content of that speech. (e.g., Arnon and Cohen Priva, 2013). Thus far, this correlation lacks a precise theory with a cognitive explanation. By producing a cognitive model of these speech rates, we provide evidence for such a theory.

2.4 Lexical Retrieval

This paper examines the time course of lexical retrieval for the case of fluent, naturalistic speech. Different facets of language can interfere with lexical retrieval in different contexts, which provides evidence toward an architecture: Schriefers et al. (1990) found that semantic, but not phonological material can cause interference, suggesting that the two are represented separately. Ratcliff and McKoon's (1989) study focuses on sentence retrieval and found that semantic information is also retrieved in stages. Here, we seek to model the retrieval process in the context of fluent speech.

There are a number of memory models in the literature that provide accounts of the timing of lexical access. For instance, classic models such as Dell's (1986)'s model of spreading activation during language production and Levelt et al.'s (1999) WEAVER++ model both provide quantitative values for retrieval times based on the form of a word.

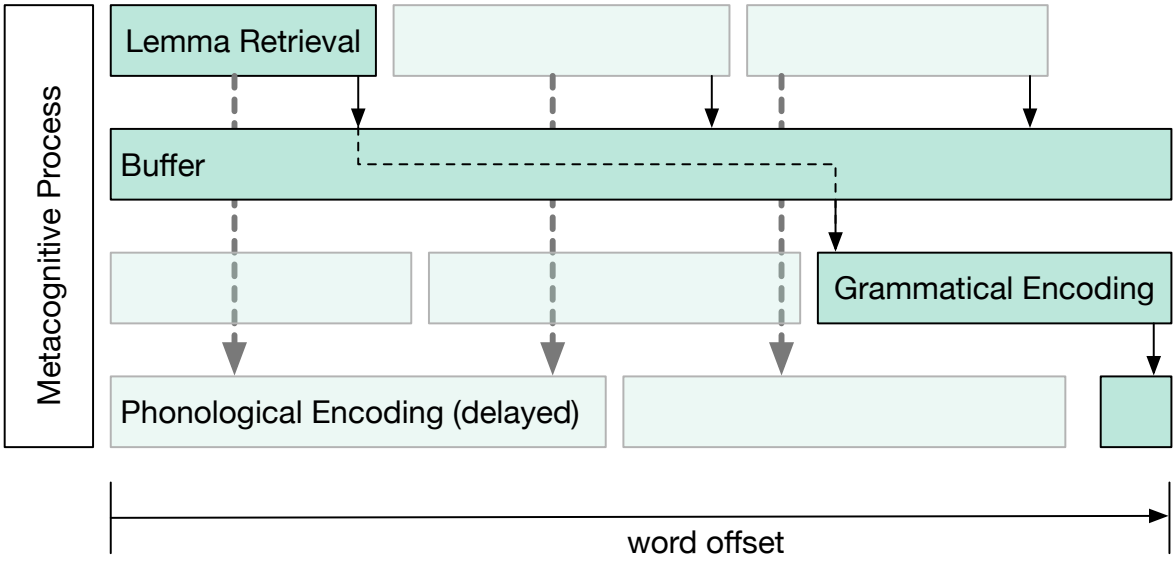

Figure 1: Our psychological model assumes that language production involves several parallel processes, and that retrieval of lemmas can interact with concurrent retrieval and/or encoding of phonological representations (dashed arrows) due to concurrent resource usage. Lemmas are retrieved several words before they are spoken. Their exact point of retrieval could depend on other factors. Likewise, while we represent phonological retrieval/encoding as a separate process for clarity, we make no claim to what extent these processes overlap.

Models such as Rapp and Goldrick (2000) focus on modeling speech errors based on word activation and context. Our model differs from these in that it attempts to model retrievals from fluent speech rates, rather than single word lexical retrieval based on picture naming tasks. Finally, while speech errors are likely related to failed lexical memory retrievals, we focus on speech that was eventually successfully retrieved and produced.

More relevantly, Dell and O'Seaghdha (1992) examine the time course of lexical access in language production. In particular, they use series of three words and EEG data to estimate lexical retrieval time. However, the lab setting it took place in precluded it as a study of naturalistic speech. Further, their model of the effects of word-properties relied on primarily qualitative attributes, such as semantic or phonetic relatedness. In particular, they find additional evidence for lemma and phonological retrieval taking place in separate stages, based on inhibition and facilitation effects. The goal of the present study is to expand the examined time frame in the hopes of replicating their argument on naturalistic speech while viewing effects found throughout, rather than just a three word window.

3 ACT-R Model

To motivate the corpus-based empirical analysis, we first describe our high-level model of the language production process. Our method primarily relies on simulating the state of lexical declarative memory during language production. After we simulate the memory retrievals for each word, we can compare this information to the actual empirical timing data in the corpus. In particular, we rely on Anderson's (1983) original account of memory. This framework was selected rather than newer or more task-specific frameworks as it is the same underlying memory model of ACT-R, which has been used to explain a wide variety of language phenomena (e.g., Vasishth and Lewis, 2004; Reitter et al., 2011), but also has been used to explain everything from decision-making (e.g., Marewski and Mehlhorn, 2011) to visual attention in graphical user interfaces (e.g., Byrne et al., 1999). Thus, by using this model, our work naturally builds upon a large body of work, using the same mechanisms to explain a variety of tasks.

Figure 1 illustrates how lemma retrieval of a target word affects phonological encoding of speaking of an earlier word. Retrieval timing is computationally estimated using the cognitive architecture ACT-R, and we assume that this retrieval

time proportionally affects phonological encoding. This can take place strategically, via a metacognitive process that coordinates these different modules, or via interference because both processes share declarative memory resources.

Our model of lexical memory is principally based on Anderson (1983)'s discussion of recency, frequency, and context effects. Activation (A) within the context of the ACT-R system is generally described by the sum of *base-level learning* (bll) and *spreading activation* (sa), which we adopt for our model as well (Anderson et al., 2004). Activation, can be defined as a linear combination of spreading activation base-level learning:

$$A(x) = sa(x) + bll(x) \tag{1}$$

For our purposes, we consider x to refer to an individual word. Base-level learning refers to the frequency and recency effects. In the base-level learning equation, it can refer to both because of the *decay* parameter, d, which causes more recent presentations to be more important, with older presentations (signified by their time of presentation, t) becoming exponentially less relevant. These older presentations, when considered together, add to the equation through their sheer quantity, providing the frequency effect, defined as:

$$bll(x) = \log\left(\sum_{i \in P_x} t_i^{-d}\right) \tag{2}$$

In this equation, 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).

$$bll(x) \approx \log\left[\sum_{i}^{k} t_i^{-d} + \frac{(n_x - k)\left(t_{n_x}^{1-d} - t_k^{1-d}\right)}{(1-d)\left(t_{n_x} - t_k\right)}\right] \tag{3}$$

While Petrov (2006) shows that the equation is close even for $k = 1$, we used $k = 5$ to more closely approximate the original equation. We then use the ACT-R default for the decay parameter, 0.5. Note that it has been suggested (e.g., Lewis and Vasishth, 2005; Cole et al., 2017) that this decay parameter could be different for language processing. In this work, we are only concerned with relative, rather than absolute values for a word's activation in memory.

In order to compute the total number of presentations, we relied on a fairly simple estimate. We multiply the number of seconds a person has been alive with the average speaking rate and that word's frequency to obtain an estimate of the amount of times a person has encountered that word; it is difficult to measure the difference between being exposed to the lexical form of the word compared to the phonological form, and it is even harder to measure any subsymbolic exposure due to thought. Still, using this formula, a unigram score computed by SRILM (Stolcke, 2002) applied to the British National Corpus, the average speaking rate of Switchboard participants (197 words/minute) as computed by Yuan et al. (2006), and the average age of Switchboard participants (37) (Godfrey et al., 1992), we can compute a baseline number of presentations for every word in Switchboard.

Next, computing spreading activation on a corpus as described in Anderson (1983) would likewise be computationally intractable. However, Pirolli et al. (2006) showed that for large sample sizes of language, Pointwise-Mutual Information is nearly identical. Therefore, we use Semilar's PMI database computed on the Wikipedia corpus (Rus et al., 2013; Church and Hanks, 1990).

In the ACT-R system, generally only items currently in working memory affect memory retrievals (Anderson et al., 2004). Likewise, we maintain the n previous words in a buffer to compute their spreading activation to the next word. We used $n = 5$ as an estimate for working memory size in language, as found in a reading task (Daneman and Carpenter, 1980). For our model, we compute the spreading activation between retrieved word, x, and each word in working memory, y, as:

$$sa(x) \approx \sum_{y}^{n} pmi(x, y) = \sum_{y}^{n} \log \frac{p(x, y)}{p(x)p(y)} \tag{4}$$

Once we have a value for activation, it's fairly simple to compute an estimate for retrieval time (RT) using the same equations from Anderson

(1983).

$$RT = I + 1/A - \frac{Ke^{-KA}}{(1 - e^{-KA})} \quad (5)$$

In this equation, I is an intercept, easily fitted with a linear model. As a parameter, K represents the cutoff time (in seconds) before there is a retrieval failure. This equation actually only represents the time required in the case of successful retrievals, which is nonetheless bounded by K, which in that sense could be thought of as the maximum possible time for a successful retrieval. While retrieval failures are part of normal ACT-R processing, they are not relevant to our model. Since our model is formed of already spoken words, they cannot represent retrieval failures. Thus, while the equation only represents successful retrievals, it is appropriate for our model. We chose the architectural default of 1.0 for K.

4 Methods

4.1 Corpus Analysis

The empirical speech data was taken from the Switchboard corpus (Godfrey et al., 1992) which is part of the Penn Treebank corpus (Marcus et al., 1993). This dataset consists of telephone conversations between strangers on a random topic, annotated to include the start and finish time for every word that has been spoken. Using our model of lexical memory as described in the previous section, we trace through the model and compute the activation of each word at its onset time.

Once the activation was computed for each word at the point when it was spoken, our goal was to observe its effect on overall speaking rates. In order to estimate when x was retrieved, we examined the speech some number of *words* back from word x. If words are spoken systematically more slowly or quickly based on word x's activation and their positional relationship to word x, then we can assume where words are spoken more slowly, retrievals are taking place. Where words are spoken more quickly, retrievals have finished. Importantly, since this is being computed at every sentence position, this should not capture positional effects. See Figure 1 for a visual depiction of our model of interference during lexical retrieval, which allows us to infer retrieval based on such interference.

While a naive model may expect lexical retrieval to occur immediately before grammatical or phonological encoding, this is not necessarily the case. Indeed, the amount of time before encoding may not be constant and may vary from word to word.

Our analysis of the corpus requires computing each word's *delay*, which is defined as the amount of time between the onsets of two sequential words, including any disfluencies that occur. As words themselves naturally can require different amounts of time to speak, we instead use the *adjusted* delay which is computed by taking the average of all of the durations of that word (as found in Switchboard) and subtracting it from the given duration. Thus, the adjusted delay could be a positive or a negative number, representing slowdowns and speedups, respectively. Throughout this paper, we use the term *delay* to actually refer to this adjusted delay. The delay referred to in Figure 1 is thus the adjusted delay: the difference between the expected delay based on the word form and the actual observed delay. To be clear, that means that if a delay term is not zero, there was a variation from the normal speed of processing, to either be quicker (negative delay) or slower (positive delay).

These speedups and slowdowns, and their relationship to retrieval time, allow us to make an argument about the interaction between lexical and phonological processing. From a statistical point of view, as we are comparing retrieval time and slowdowns in the same units, our linear model could be thought of as the percentage of retrieval time that is behaviorally reflected in language production.

4.2 Experiment

Data were analyzed with two related models. Initially, we tested an *interaction* model in order to test our hypothesis of the interaction between delay and offset (see Table 1). From this information, we use exploratory data analysis in the form of a *discrete* model, in order to explore the critical regions of the graph (see Table 2). From this exploratory data analysis, we present the pooled version of the discrete model for easier interpretation of our found effects (see Table 3). For both models, the activation of a target word and its expected retrieval time burden was computed, as were the delays for the n words preceding the target word. Importantly, note that in both models, when we refer to the expected retrieval time or activation, we

are referring to the target word, not any of the preceding words. Both models are concerned with the word *offset* (*i*), which refers to the number of interceding words between the given delay and the target word, such that $\mathrm{i} = 0$ refers to the word immediately before the target word.

In the interaction model, we are interested in the interaction term between word offset and delay: its goal is to show how the correlation changes with offset. In this model, every observation only uses a single offset, chosen randomly, for each target word. All of the other observations for that word are discarded. This is to ensure the observations are independent. The correlation coefficients of interest are the correlation of delay as a whole, and its interaction effect with offset. In general, the coefficient of offset by itself is likely capturing some distributional information about the data, rather than anything interesting with how it relates to memory retrievals. As a linear model:

$$RT \sim \text{delay} * \text{offset}$$

Meanwhile, the discrete model's observations consist of a word's expected retrieval time and the delays from previous words. Then, we make a linear model using each of the delays as a predictor. Note that in this notation, delay_i refers to the delay of offset word i. To reiterate, i represents how many interceding words there are between that offset word and the target word. As a linear model, this would be:

$$RT \sim \text{delay}_0 + \text{delay}_1 + ... + \text{delay}_n$$

The goal of the interaction model is to show the robustness of the slope associated with index, while the goal of the discrete model is to allow for a non-linear relationship between offset and the effect of delay on activation, examining up to 25 previous words. Exploring this non-linear relationship allowed us to infer the critical regions of this effect. Importantly, the discrete model's goal was to explore the significant relationship found in the interaction model more deeply, rather than to itself justify the effect.

Under the model shown in Figure 1, we expect that longer retrieval times of the target word are associated with slowdowns of speech production at some time before the target word is spoken. Earlier than that point, the target word should have no influence on speech production.

	Estimate	Std. Error	t-value	p-value
(Intercept)	.1796	.0002	728.715	$<$ .00001 ***
offset	.0015	.0010	1.1512	.011 **
delay	.0668	.0048	13.839	$<$ .00001 ***
delay*offset	-.0066	.0005	-12.411	$<$ .00001 ***

F-stat	DF	p-value	Adj R^2	Multi R^2
102.2	802055	$<$ .00001	0.0005	0.0005

Table 1: Linear regression predicting expected retrieval time of a target word as a function of the delay in speaking of a previous word at that offset.

	Estimate	Std. Error
I	.1844	.0002
d0	-.0033	.0001
d1	-.0011	.0001
d2	-.0006	.0001
d3	-.0002	.0001
d4	-.0001	.0001
d5	-.0001	.0001
d6	.0002	.0001
d7	.0001	.0001
d8	.0002	.0001
d9	-.0001	.0001
d10	.0002	.0001
d11	.0000	.0001
d12	.0002	.0001
d13	.0000	.0001
d14..d24	0.000.	0.000

Table 2: The linear effects model relating each discrete delay term with expected retrieval time. A higher number on the delay term signifies the number of words between the delayed word and the target word. This exploratory data analysis was done to inform the pooled model. Also see Figure 2.

5 Results and Discussion

If one focuses on the interaction model, our experiments yield a relatively counterintuitive result: namely, delay is correlated in the direction opposite to what is expected. One would imagine that delay and retrieval time should be positively correlated: if people are speaking words more slowly (positive delay), then likewise, their retrieval time should be higher. However, we discovered a robust effect in the opposite direction: higher delays imply shorter expected retrieval times, and shorter delays imply longer expected retrieval times. In other words, when people are expected to need the longest to retrieve words, they actually speak more quickly, and vice versa.

	Estimate	Std. Error	t-value	p-value	
(Intercept)	.1843	.0002	728.715	$< .00001$	***
early	-.0052	.0010	-46.36	$< .00001$	***
early(ns)	0.000	0.000	-0.8200	.412	
late	.0009	.0002	5.448	$< .00001$	***
late(ns)	.0002	.0002	1.263	.2070	

F-stat	DF	p-value	Adj R^2	Multi R^2
567.9	777924	$< .00001$	0.003	.003

Table 3: Pooled version of discrete linear model, based on critical regions from the graph. Regions are broken at 3, 5, and 14 respectively.

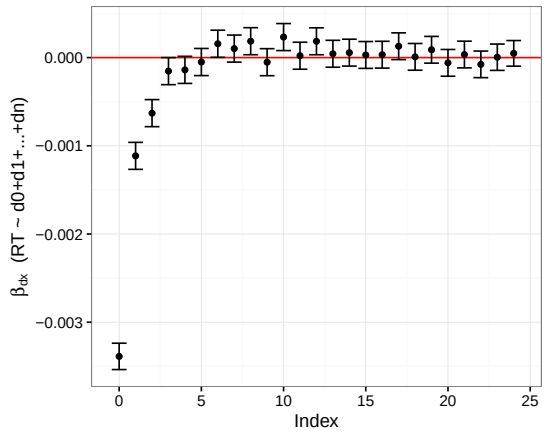

Figure 2: The discrete model's linear predictors (see also Table 2). Error bars represent normal 95% confidence intervals. This graph and Figure 3 have similar critical regions, which informed the pooled model presented in Table 3: 0-2 appear significant and negative, 3-4 are not significant (slightly negative), 5-14 are significant and positive, 15+ is not significant.

Examining the effect for larger offsets, however, we observe that the effect reverses before disappearing. Thus, we see an effect in the expected direction for the delays of word offsets 4 through 14. This is commensurate with word planning that takes place several words in advance rather than immediately before the word; likewise, the effect also disappears in the interaction model based on the interaction effect.

See also Figure 2 and Figure 3, which are visualizations of the discrete and interaction model, respectively. These graphs show how the relationship between activation of a word and speech delay develops over the offsets, i, before the word. While Figure 2 has its effects pulled directly from Table 2, Figure 3 is produced by raw data, defined by:

$$y(j,i) = \frac{A_j - \beta_0}{\text{delay}_i} \tag{6}$$

These graphs were designed to demonstrate how the effect switches from positive to negative as we move back from immediately before the word to earlier in the utterance. With the interaction model, we wanted to show statistical evidence for the pattern of effects; the discrete model quantifies the gradual fade to zero. We interpret the models as follows.

1. There is a strong negative correlation of the word delays with expected retrieval time for the words immediately before the target word. Since retrieval time is a function of activation, this would imply that the observable phonological effect happens later for more activated words, which are likely retrieved shortly before their use.
2. There is a weaker but significant positive correlation of the word delays with expected retrieval time for words about 5-14 words preceding the target word. These delays likely occur for words with less activation, whose retrievals are likely initiated early to ensure that there is enough time.
3. For words very far away from the target word, there is no reliable effect, implying that this is not just an effect of a cyclical information distribution.

6 General Discussion

These results confirm some classical findings on lexical retrieval, while adding a subtle but reliable new effect. Further, these findings have some implications for incrementality and uniform information density.

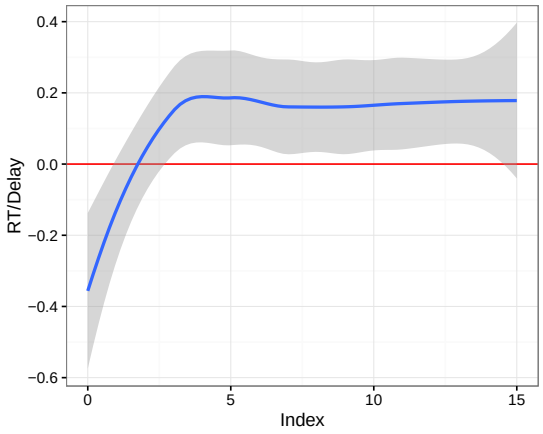

Figure 3: Smoothed correlation between delay and expected retrieval time across offset, created using a sample of raw data (representative of the interaction model). Effect disappears after offset 15, but full graph is not shown to avoid smoothing small but significant effects with non-significant effects.

In our discussion, we will frequently refer to the *activation* of a word. Recall that activation in the ACT-R sense is the inverse of the expected retrieval time: higher activation implies a shorter expected retrieval time. While retrieval time makes more sense in a time-predictive linear model, it is easier to interpret our results based on its relationship to activation.

6.1 Lexical Retrieval

It is difficult to separate the lexical retrieval effects we found into the two categories of retrievals described by Levelt (1992): a lemma retrieval and a later phonological retrieval. However, this is not to claim that they cannot be, but simply that our methodology did not easily allow us to. A commonly implied assumption is that lemma retrievals shouldn't interfere with phonological processes (e.g., Schriefers et al., 1990), though it is difficult to know if a speech slowdown is due to a phonological or semantic interference due to our experimental setup. However, since in our experiment, effects are still observed at large distances from the target words, either phonological forms can be retrieved in a non-incremental way (possibly even before lemmas for other words are retrieved), or the retrieval of the lemma does interfere with phonological encoding in some way; for instance, by activating related phonological forms. Still, we ultimately find the same pattern of effects as Dell and O'Seaghdha (1992): facilitatory effects close to the target word, with inhibitory effects further away. The primary difference is the time frame, which is possibly due to their experimental setup.

6.2 Process Model

We found a surprising effect: words with higher activation are not spoken more quickly, but more slowly. This also applies to the words that immediately precede them. However, if we look further back, we see a robust effect in the expected direction: if the approaching word has a high activation, they are said more quickly, but if the approaching word has a low activation, they are said more slowly. We argue that this slowdown is the result of shared resources between phonological and grammatical encoding, and as activation directly predicts retrieval time, we posit that word retrievals are part of what causes slowdowns. The corresponding speedups could be because the work of planning the sentence up to that point is then done. The most important prediction of this is that it means low activation words are retrieved earlier, which would imply that there is some cognitive strategy facilitating the necessity of initiating early retrievals for low activation words.

6.3 Incrementality

These results provide information about the timing of memory retrievals, given that such retrievals are related to activation. As activation is inherently related with how long a memory retrieval should take, it makes sense there are some cognitive strategies for coping with this disparity in order to produce seemingly fluent dialogue. That strategy involves buffering: retrieving and storing the words that will need longer to retrieve, based on the structure of the sentence.

Further, this type of buffering strategy could be part of the strategy that Ferreira and Swets (2002) refer to, when they propose the incrementality of language production is under "strategic control." While a purely incremental strategy might have interlocutors retrieve in a purely incremental fashion, there are some hiccups: certain words take longer to retrieve than others. By this logic, if grammatical encoding proceeds in a purely incremental fashion, then lexical retrieval does not, and vice versa. Thus, it is reasonable to believe that the grading of incrementality found in natural human discourse is not only variable from situation to situation, but it may be variable amongst competing processes for any given situation.

6.4 Uniform Information Density

Let's consider an additional explanation. The *Constant Entropy Rate* Hypothesis (Genzel and Charniak, 2002) posits that lexical material is distributed across a sentence (and other units) such that its information is held approximately constant. Could a difficult-to-retrieve, slow word at position j be likely to be combined with easier-to-retrieve, high-frequency words at positions $j-4...j-1$, causing the significantly increased speech rate we found there?

The model of buffered retrievals, along with the empirical evidence, may provide a cognitive mechanism that results in an approximately constant entropy rate. Thus, Uniform Information Density (UID, e.g., Jaeger, 2010) could be considered a consequence of the cognitive procedures involved in retrieving syntactic-lexical items from

declarative memory while grammatically encoding those materials retrieved earlier.

7 Future Work

Our work opens up several possible avenues for future research. While it is unclear if syntax rules are retrieved from some form of implicit memory (e.g., Reitter et al., 2011), lexical items clearly are. Syntactic processing could potentially adapt to working memory, rather than itself guide lexical retrievals (e.g., Cole and Reitter, 2017). By this argument, memory retrieval is a largely automatic, rather than attention-driven process, and syntax makes use of what is available to produce fluent dialogue. In this type of model, the constant size of the retrieval buffer would provide a clear corollary to Uniform Information Density.

Furthermore, this paper does not clearly differentiate between lemma and phonological retrieval. Although we do not expect phonological forms to be retrieved as early as the effects we are seeing, we also do not expect lemma retrieval to have effects on phonological encoding. A computationally implemented process model could explore these effects in more detail.

Lastly, this study provides another mechanism by which non-sequential dependencies in language production are observed. It seems possible that non-incremental language processing can be explained as a process that involves general memory mechanisms including cue-based memory retrieval. What is in question is whether we really process local syntax using structured, memory-hungry models (i.e., with syntax trees); we note that in natural language processing, skip-grams can capture local, non-incremental relationships among words. Thus, the relationship between working memory, syntax trees, and skip-grams appears to be of continued interest.

8 Conclusions

In this paper, we explore the process of lexical memory retrieval in the context of language production. In contrast to previous work, we look at a corpus of natural speech and do not rely on single word retrievals in an experimental setting. This allows us to observe how certain processes involved in fluent language production overlap. In particular, the data support a model according to which lexical retrievals can happen quite early. By using the formalism defined by the empirically-validated ACT-R framework, we show when memory retrievals are taking place through the effect on speaking rates, seeing facilitation early and inhibition later. We conclude that low-activation words can be retrieved as early as 14 words before they are spoken. As low activation words are higher information and require longer to retrieve, this has theoretical implications for some empirical findings of language processing.

Acknowledgements

This project was supported by National Science Foundation projects BCS-1457992 and IIS-1459300. We would like to thank Alex Ororbia, Matthew Kelly, Yang Xu, and Ying Xu for their comments on an earlier version of the paper.

References

John Anderson. 1983. A Spreading Activation Theory of Memory. *Journal of Verbal Learning and Verbal Behavior* 22:261–295.

John R. Anderson. 1991. Cognitive architectures in a rational analysis. In Kurt VanLehn, editor, *Architectures for Intelligence*, Lawrence Erlbaum Associates.

John R. Anderson, Daniel Bothell, Michael D. Byrne, Scott Douglass, Christian Lebiere, and Yulin Quin. 2004. An integrated theory of the mind. *Psychological Review* 111:1036–1060.

Inbal Arnon and Uriel Cohen Priva. 2013. More than words: The effect of multi-word frequency and constituency on phonetic duration. *Language and Speech* 56(3):349–371.

Inbal Arnon and Uriel Cohen Priva. 2014. Time and again: The changing effect of word and multiword frequency on phonetic duration for highly frequent sequences. *The Mental Lexicon* 9(3):377–400.

Alan Bell, Jason M Brenier, Michelle Gregory, Cynthia Girand, and Dan Jurafsky. 2009. Predictability effects on durations of content and function words in conversational english. *Journal of Memory and Language* 60(1):92–111.

J Kathryn Bock and Willem J M Levelt. 2002. *Language production: Grammatical encoding*, Routledge, volume 5, pages 405–452.

Michael D Byrne, John R Anderson, Scott Douglass, and Michael Matessa. 1999. Eye tracking the visual search of click-down menus. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, pages 402–409.

Alfonso Caramazza. 1997. How many levels of processing are there in lexical access? *Cognitive Neuropsychology* 14(1):177–208.

Alfonso Caramazza. 2006. Lexical access in bilingual speakers: What's the (hard) problem? *Bilingualism: Language and Cognition* 9:153–166.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1):22–29.

Jeremy R. Cole, Moojan Ghafurian, and David Reitter. 2017. Linking memory activation and word adoption in social language use via rational analysis. In *Proceedings of the 15th International Conference on Cognitive Modeling*. London, UK.

Jeremy R Cole and David Reitter. 2017. Examining Working Memory during Sentence Construction with an ACT-R Model of Grammatical Encoding. In *Proceedings of the 15th International Conference on Cognitive Modeling*. London, UK.

Allan M Collins and Elizabeth F Loftus. 1975. A spreading-activation theory of semantic processing. *Psychological Review* 82(6):407.

Meredyth Daneman and Patricia A Carpenter. 1980. Individual differences in working memory and reading. *Journal of verbal learning and verbal behavior* 19(4):450–466.

Gary S. Dell. 1986. A spreading-activation theory of retrieval in sentence production. *Psychological Review* 93(3):283.

Gary S. Dell and Padraig G. O'Seaghdha. 1992. Stages of lexical access in language production. *Cognition* 42(1):287–314.

Gary S. Dell, Myrna F. Schwartz, Nadine Martin, Eleanor M. Saffran, and Deborah A. Gagnon. 1997. Lexical access in aphasic and nonaphasic speakers. *Psychological Review* 104(4):801–838.

Fernanda Ferreira and Benjamin Swets. 2002. How incremental is language production? Evidence from the production of utterances requiring the computation of arithmetic sums. *Journal of Memory and Language* 46(1):57–84.

Victor S. Ferreira. 1996. Is it better to give than to donate? Syntactic flexibility in language production. *Journal of Memory and Language* 35:724–755.

Victor S. Ferreira and L. Robert Slevc. 2007. Grammatical encoding. In M. Gareth Gaskell, editor, *The Oxford Handbook of Psycholinguistics*, Oxford University Press, page 453.

Dmitriy Genzel and Eugene Charniak. 2002. Entropy rate constancy in text. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pages 199–206. https://doi.org/10.3115/1073083.1073117.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. SWITCHBOARD: telephone speech corpus for research and development. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-92)*. IEEE, volume 1, pages 517–520.

Markus Guhe. 2007. *Incremental Conceptualization for Language Production*. Lawrence Erlbaum Associates.

T Florian Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology* 61(1):23–62.

Otto Jespersen. 1992. *The philosophy of grammar*. University of Chicago Press.

Willem JM Levelt. 1992. Accessing words in speech production: Stages, processes and representations. *Cognition* 42(1):1–22.

Willem JM Levelt, Ardi Roelofs, and Antje S Meyer. 1999. A theory of lexical access in speech production. *Behavioral and brain sciences* 22(1):1–38.

Richard L. Lewis and Shravan Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive science* 29(3):375–419.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *COMPUTATIONAL LINGUISTICS* 19(2):313–330.

Julian N Marewski and Katja Mehlhorn. 2011. Using the ACT-R architecture to specify 39 quantitative process models of decision making. *Judgment and Decision Making* 6(6):439.

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

Martin J. Pickering and Simon Garrod. 2013. An integrated theory of language production and comprehension. *Behavioral and Brain Sciences* 36(04):329–347.

Peter Pirolli, Wai-tat Fu, Ed Chi, and Ayman Farahat. 2006. Information scent and web navigation: Theory, models and automated usability evaluation. *The Next Wave: NSA's Review of Emerging Technologies* 15(2):5–12.

Brenda Rapp and Matthew Goldrick. 2000. Discreteness and interactivity in spoken word production. *Psychological review* 107(3):460.

Roger Ratcliff and Gail McKoon. 1989. Similarity information versus relational information: Differences in the time course of retrieval. *Cognitive Psychology* 21(2):139–155.

David Reitter, Frank Keller, and Johanna D Moore. 2011. A computational cognitive model of syntactic priming. *Cognitive Science* 35(4):587–637.

Ardi Roelofs, Antje S Meyer, and Willem J M Levelt. 1998. A case for the lemma/lexeme distinction in models of speaking: Comment on caramazza and miozzo (1997). *Cognition* 69(2):219–230.

Vasile Rus, Mihai C Lintean, Rajendra Banjade, Nobal B Niraula, and Dan Stefanescu. 2013. Semilar: The semantic similarity toolkit. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria, pages 163–168.

Herbert Schriefers, Antje S. Meyer, and Willem J.M. Levelt. 1990. Exploring the time course of lexical access in language production: Picture-word interference studies. *Journal of Memory and Language* 29(1):86–102.

Andreas Stolcke. 2002. SRILM-an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*. Denver, volume 2, pages 901–904.

Shravan Vasishth and Richard L. Lewis. 2004. Modeling sentence processing in ACT-R. In *Proceedings of the Workshop on Incremental Parsing: Bringing Engineering and Cognition Together*. Association for Computational Linguistics, pages 82–87.

Jiahong Yuan, Mark Liberman, and Christopher Cieri. 2006. Towards an integrated understanding of speaking rate in conversation. In *Proceedings of the 9th International Conference on Spoken Language Processing*. Pittsburgh, Pennsylvania.