On Factors Influencing Typing Time: Insights from a Viral Online Typing Game

Robert Chen, Roger Levy & Tiwalayo Eisape
{robertcc,rplevy,eisape}@mit.edu
Brain and Cognitive Sciences
Massachusetts Institute of Technology

Abstract

Context effects in human spoken language are well-documented and play a central role in the theory of language production. However, the role of context in written language production is far less well understood, even though a considerable proportion of the language produced by many people today is written. Here we analyze the factors predictive of English language typing times in a large, naturalistic corpus from the popular *TypeRacer.com* website. We find broad consistency with the major documented effects of linguistic context on spoken language production, suggesting potential modality-independence in the cognitive mechanisms underlying language production and/or similar optimization pressures on the production systems in both modalities.

Keywords: Psycholinguistics; Language production; Typing speed; Prediction; Context effects; Planning

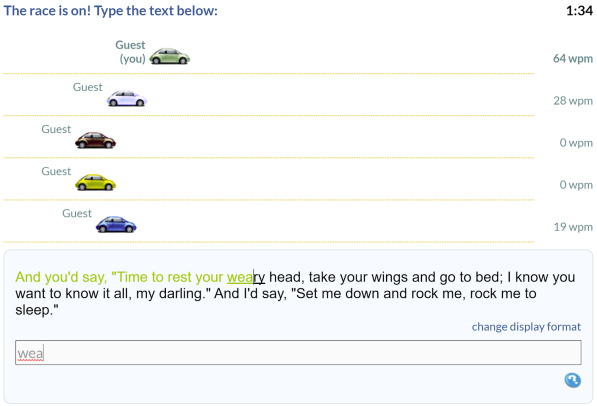

Figure 1: An in-progress race on TypeRacer.com between 5 "guest" players. Typing time in words per minute and the position of each player's car icon are continuously updated according to the players' typing speed and relative position in the race text.

Introduction

Language production is work: formulating messages, retrieving the corresponding phonological or orthographic forms, and executing the appropriate motor actions to express them are effortful, take time, involve delays, and are error-prone (Levelt, 1993). This difficulty varies in fine-grained temporal scale within an utterance: for example, disfluencies are more likely early in phrases than later (Boomer, 1965; Shriberg, 2001). One source of this variability in difficulty has to do with the unit being produced at any given moment: words that are high in frequency and phonotactic probability are produced faster in naming studies (Oldfield & Wingfield, 1965; Balota & Chumbley, 1985; Vitevitch, Armbrüster, & Chu, 2004); words with high frequency and low average information content have shorter durations in continuous spoken language (Umeda, 1977; Pluymaekers, Ernestus, & Baayen, 2005; Seyfarth, 2014). A word's form similarity to other words in the lexicon also affects its articulation, though the nature of that relationship remains more controversial (Gahl, Yao, & Johnson, 2012; Scarborough, 2012; Meinhardt, Baković, & Bergen, 2020).

But another key source of variability is the linguistic *context* in which the unit is produced. In particular, words that are more predictable in their context are shorter in duration, have less strongly articulated segments, and are less prone to disfluency (Shriberg & Stolcke, 1996; Aylett & Turk, 2006; Bell, Brenier, Gregory, Girand, & Jurafsky, 2009; Arnon & Cohen Priva, 2013). This relationship between a word's probability in context and the details of its form realization suggests optimization in the mechanisms of human language production: the more a word is predictable from, and thus redundant with, its context, the less time, effort, and detail is necessary to realize it (Jurafsky, Bell, Gregory, & Raymond, 2001; Levy & Jaeger, 2007). However, the degree to which this optimization is purely speaker-centric versus listener-oriented remains unresolved (Arnold, 2008; Watson, Arnold, & Tanenhaus, 2008; Galati & Brennan, 2010; Hall, Hume, Jaeger, & Wedel, 2018).

Today, over 86% of the world over age 15 is literate (Roser & Ortiz-Ospina, 2016). Although we are not aware of published statistics on the matter, a considerable proportion of this population may produce written language in quantity comparable to spoken language. Furthermore, the ability to produce written language quickly and accurately is of con-

siderable value in the ability to thrive in modern literate societies. Since written language production is an important and highly practiced activity for this population, it might plausibly be under the same optimization pressures as spoken language production. A clear empirical picture of the relationship between lexical and contextual features and the patterns of written language production could improve our understanding of human language production more generally. Yet our empirical understanding of this relationship is considerably less developed than for spoken language.

Here we attempt to help clarify this empirical picture, through the study of one mode of written language production: typing. Probability effects in typed language are well-documented at the lexical and sublexical level (Terzuolo & Viviani, 1980; Gentner, 1982; Gentner, Larochelle, & Grudin, 1988; Weingarten, Nottbusch, & Will, 2004; Cohen Priva, 2010), consistent with the view that word-specific motor production routines are stored and optimized through experience. However, little is known about the effects of context beyond the individual word, presumably because studying these effects require larger datasets and more advanced modeling techniques to effectively estimate word probability in context and its effect on typing behavior.

Here we address these limitations by taking advantage of a very large dataset derived from publicly available data on the *TypeRacer.com* website, where users compete against each other to correctly type short website-provided texts as quickly as possible. Coupled with contemporary models from natural language processing, this dataset allows us to investigate a wide range of determinants of typing behavior in light of contemporary theories of language production. Overall we find similar general shape of effects of word properties and context-based predictability on typing time as has been documented for word duration in spoken language production, but we also find key differences in the detailed relationships.

TypeRacer

More so than in spoken language production, typing proficiency is a function of speed (among other factors e.g., error rate). Typists worldwide dedicate significant time to honing this skill and, in the process, generate massive amounts of data. These data offer a potential alternative to spoken language corpora in shedding light on human language production.

TypeRacer is a viral online typing game where players race against themselves, friends, or strangers in groups of up to 10 to type a short text as quickly as possible (Fig. 1). Typists are given up to 12 seconds to read the race prompt before the race starts. Gamification of this sort not only allows TypeRacer to collect data from thousands of players across the world, resulting in a massive and multilingual dataset, several orders of magnitude larger than what can be collected in the lab. It also motivates players to type efficiently, in order to win. As we will show, a considerable amount of typing time variability in TypeRacer can be attributed to linguistic factors theoretically tied to processing effort and unlikely to be confounded with motivation. This makes the dataset of considerable interest for studying cognitive and motor constraints in language production. TypeRacer has races in 50 languages and includes diverse text such as song lyrics and code making it a diverse, openly accessible dataset with data from over $N = 100,000$ users, across $35,000$ distinct texts, and over 7×10^7 individual races.

Dataset

We restrict our analysis to a random subset of the data on the TypeRacer website consisting of 1000 English-typing [1] TypeRacer users. We call our dataset TypeRacer1000 [2]. For each user we acquire typing data from up to 100 randomly sampled races (without replacement). Only races completed after June 3, 2017 were considered [3]. Data for each race on TypeRacer consists of the ordered list of keystrokes along with the time, in milliseconds, of each keystroke (i.e., time since last keystroke, or start of race for the first keystroke). We calculate word typing time as the sum of the keystroke times of all keystrokes of the word. TypeRacer1000 has a total of 60 thousand races and 3.0 million measures of word typing times ($\mu = 49.9$ words per race, $\sigma = 20.3$).

Predictors

Language Model Surprisal We use autoregressive language models (LMs) trained on Wikitext-2 (Merity, Xiong, Bradbury, & Socher, 2016) to estimate in-context probability for the words in our dataset (see Dataset Details for training details). Because the complete text prompt is provided to users before the race starts, users may plausibly encode not only the left context of a particular word in the prompt for use in planning that word's typing, but also the right context of that particular word. This is an interesting analogue to the case of naturalistic speech production, where in general a speaker may have done some planning and encoding of a message representation corresponding to the right context of the word being uttered at any moment (Momma, Slevc, & Phillips, 2016), and indeed predictability-based speech modulation effects from right context seem stronger than from left context (Bell et al., 2009). To investigate the respective roles of left and right context on typing, we estimate our models from Wikitext-2 in both the forward direction (i.e., typical autoregressive langauge modeling based on **Left** context) and in the backward direction (**Right** context). At both training and inference time, each model is given full context from the beginning or end of the prompt.

Because we seek to characterize the precise shape and size of the effect of word predictability on typing time, we draw on past work that relates predictability and human reading time (Smith & Levy, 2013; Luke & Christianson, 2016; Wilcox,

[1] English typers were defined as users with only English typing data recorded.

[2] We make our dataset available at `https://osf.io/d3z8v/`

[3] Before this date only WPM and average accuracy were available, not the full typing log.

Predictor	Description
Orthographic Neighbors	The number of orthographic neighbors (single letter substitution, deletion, or addition).
Dominance	A measure of the degree of control one feels (dominant, controlled).
Semantic Size	A measure of magnitude (big, small) expressed in either concrete or abstract terms.
Concreteness	A measure of the degree to which something can be experienced by our senses.
Repeats	The number of times the word has been repeated in the context.
Familiarity	A measure of a word's subjective experience (familiar, unfamiliar).
Imageability	A measure of the degree of effort involved in generating a mental image of something.
Valence	A measure of value or worth (positive, negative).
Arousal	A measure of internal activation (excitement, calmness).
Gender Association	A measure of the degree to which words are considered to be masculine or feminine.
Age of Acquisition	A measure of the age at which adults estimate they first learned a word.

Table 1: Descriptions of non-surprisal predictors. We use the number of orthographic neighbors, number of repeats of the word, and the Glasgow norms, which are determined via human ratings. Arousal, Valence, and Dominance are measured on 9-point scales, while the remaining Glasgow norms are measured on 7-point scales. Descriptions of Glasgow norms are copied from (Scott et al., 2019).

Gauthier, Hu, Qian, & Levy, 2020; Eisape, Zaslavsky, & Levy, 2020) and use word *surprisal* as our dependent measure. We transform the probability placed on a given word at position i, i.e., x_i, given its preceding context (or subsequent context in the case of our **Right** context model) $\mathbf{x}_{<i}$ into surprisal by taking the negative log of that probability for each of our language models:

$$S_i = -\log_2 P_{\text{model}}(x_i|\mathbf{x}_{<i}), \quad (1)$$

Similarly, we compute unigram surprisal by taking the log transform of word frequency estimates from the `wordfreq` Python library (Speer, Chin, Lin, Jewett, & Nathan, 2018) in the same fashion[4].

Language Model Training Surprisal values were estimated with 2-layer long short-term recurrent neural networks (Hochreiter & Schmidhuber, 1997) with 400 hidden units per layer and 400 dimensional word embeddings, written in PyTorch (Paszke et al., 2019) and trained via stochastic gradient descent for 40 epochs. Additional hyperparameters: initial learning rate = 20, batch size = 20, and batch length = 35.

Other Predictors In addition to surprisal estimates, we included several factors with attested effects on language processing and production in other modalities. For coverage, we include (log transformed) orthographic neighborhood density estimates from the CLEARPOND database (Marian, Bartolotti, Chabal, & Shook, 2012), several of the Glasgow norms (Scott et al., 2019), and the number of preceding repetitions of the word in the prompt (from the left) up to the point it is encountered (Table 1).

Dataset Details

Preprocessing We use the Moses tokenizer (Koehn et al., 2007) to split text on TypeRacer into words. We further modified Moses to combine tokens beginning with an apostrophe followed by one or more characters with the previous token. Because our word tokenization differs slightly from WikiText, some words in our dataset consist of multiple WikiText tokens. In these cases, we compute the surprisal as the sum of the surprisals of the tokens that make up the word.

Demographics Of our 1000-user sample, 50 do not report location, 432 are in the United States, followed by 73 in Canada, 54 in India, 51 in the United Kingdom, and 340 from other countries. 946 do not report age; the ages of the remaining range from 13 to 46 ($\mu = 23.1, \sigma = 6.79$), with one outlier (120). 783 do not list their keyboard layout, 203 use Qwerty, and 14 list other keyboard layouts. More demographic information can be found at `https://osf.io/r8z9j/`.

Results

For the results presented here we include only typists who typed at least 95 distinct texts, resulting in 462 distinct users and $N = 2.3$ million overall (word, context) pairs.

The Effect of Surprisal on Typing Time is Linear

We use generalized additive models (GAMs; Wood (2006)) to determine the functional form of major word predictors available for all words in our dataset — surprisal based on **Left** context, **Right**, and raw frequency (unigram surprisal). Because effects of these factors vary between participants (and between words) it is essential to account for these differences with by-participant and by-word offsets in our models (i.e., the "maximal" random effects structure; Barr, Levy, Scheepers, and Tily (2013)). However, the "maximal" random effects structure can be computationally challenging to use in the context of GAM fits for situations where the random effects are crossed. As a surrogate, we follow Smith and Levy (2013) in using a hierarchical bootstrap—first resampling from individual clusters (users or word types) and then by individual observations within each cluster $B = 500$ times. We then use the mean and [0.025,0.975] quantiles of the bootstrap-replicate estimates as the effect estimate and 95% confidence interval (Fig. 2). This does not give us a sin-

[4]To avoid word counts of 0, which would yield infinite surprisal, we pad out-of-vocabulary items in this model with a value of 10^{-8}

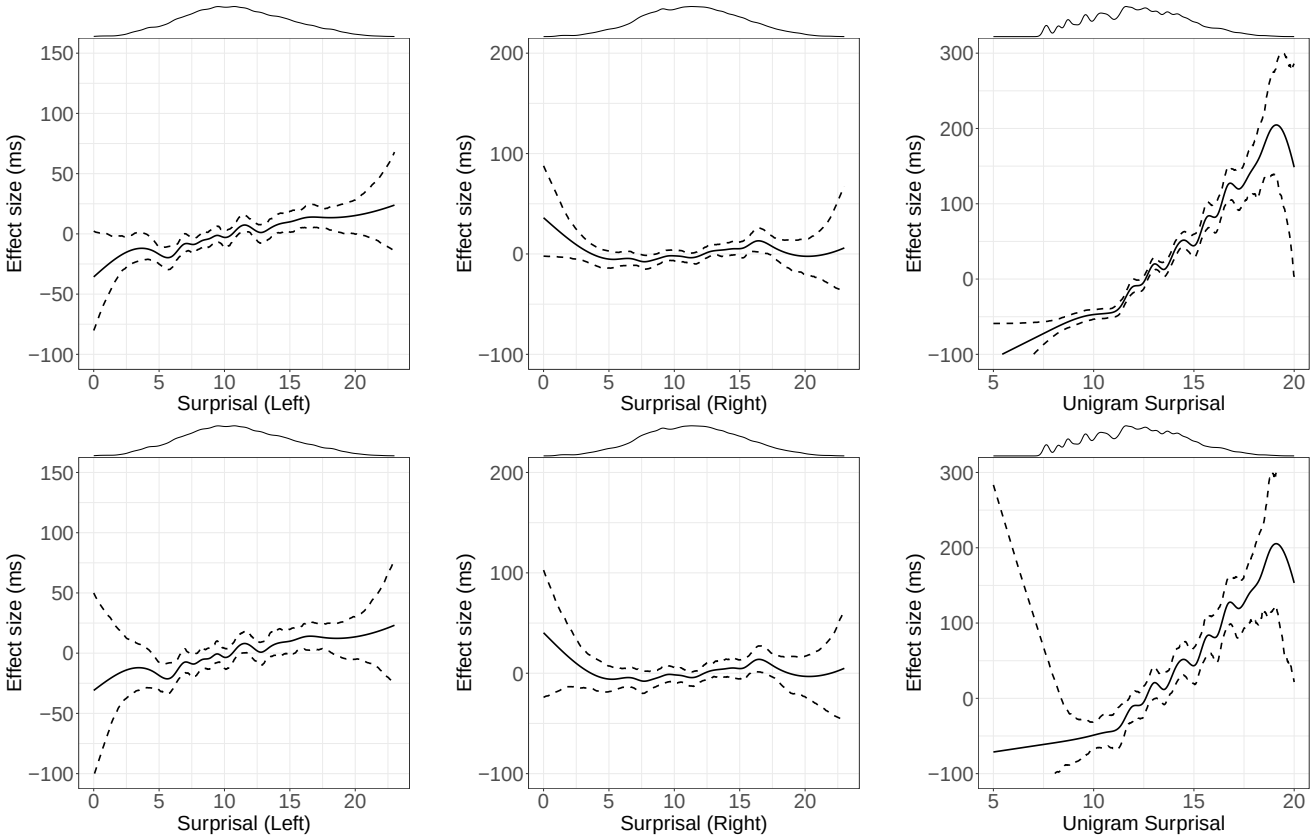

Figure 2: Relationship between various estimates of contextual probability on typing speed slowdown. The top row shows the recovered effect shape of log transformed predictors (surprisal values) using a GAM with by-User hierarchical bootstrapping, the bottom row shows the model estimated with by-word bootstrapping. Regression lines from GAM models fitted via hierarchical bootstrap are shown as solid lines 95% confidence intervals are shown as dashed lines. The marginal density of each predictor is shown at the top of each plot.

gle analysis that takes into account the full crossed repeated-measures structure of our dataset, but consistency in the results from both by-users and by-words hierarchical bootstrapping would provide evidence as to the likely underlying the functional form (shape) of these predictor–typing time relationships (Fig. 2). Because our GAMs do not make a linearity assumption, if the recovered effect shape is well approximated as linear (as is the case in e.g., reading times (Smith & Levy, 2013)), it would license treating them as such in a single mixed regression model that fully accounts for the crossed repeated-measures structure of our data.

Indeed, in the predictor ranges where most of the data lie, typing times are roughly linear in word length (not shown), unigram surprisal, forward surprisal, and backward surprisal (Fig. 2). We therefore treat these relationships as linear in the analyses we report in subsequent sections.

Prediction vs. Planning in Typing

Work in spoken language production has found that speaker *planning*, as quantified with word likelihood conditioned on right context, had a larger effect on word duration than left context predictability (Bell et al., 2009). We use two different types of mixed modeling approach and show quantitatively similar results. In both approaches, we use the surprisal, frequency, and word length predictors from the previous analysis as well as the full range of predictors we include from CLEARPOND and the Glasgow norms. The first approach is a single-stage mixed linear model with the maximal random effects structure, estimated on a random subset of 100 typists (out of the 462 in the larger dataset, subsampling for model fitting speed). The second is a "two-stage" (Gelman, 2005) linear mixed effects model: in stage one, we fit a separate model for each participant, using by-word random effects; in stage 2, we examine the distribution of the recovered model parameters for each of the 462 typists in our dataset, estimating their effect size with the distribution mean and assessing significance via t tests (Table 2).

The estimated effects sizes for in-context surprisal predictors are shown in the top portion of Table 2. We show that both surprisal effects are significant for our two-stage model (for our single stage model, only **Left** context is significant), **Left** context effects are manifestly larger than **Right** context

Predictor	1-stage Estimate	1-stage Std. Error	1-stage p	2-stage Estimate	2-stage Std. Error	2-stage p	Predictor units
Right context Surprisal	0.61	0.76	0.2	0.69	0.27	0.010	bits
Left context Surprisal	2.09	0.59	< 0.001	2.03	0.25	< 0.001	bits
Unigram Surprisal	25.16	3.28	< 0.001	19.24	0.81	< 0.001	bits
Orthographic Neighbors	-44.47	3.72	< 0.001	-36.07	1.31	< 0.001	log(neighbors)
Dominance	-7.04	5.09	0.9	-7.22	1.30	< 0.001	9-pt Likert scale
Semantic size	-1.25	4.43	0.3	-6.69	1.05	< 0.001	7-pt Likert scale
Concreteness	-7.29	5.97	0.8	-3.42	1.57	0.030	7-pt Likert scale
Repeats	-0.70	2.92	0.5	-2.55	1.66	0.125	repeats
Familiarity	8.71	8.57	0.1	3.01	2.23	0.178	7-pt Likert scale
Imageability	7.05	6.20	0.1	4.57	1.65	0.005	7-pt Likert scale
Valence	5.59	3.44	0.05	3.60	0.95	0.001	9-pt Likert scale
Arousal	6.59	4.37	0.06	8.89	1.17	< 0.001	9-pt Likert scale
Gender Association	14.90	4.95	0.001	10.99	1.36	< 0.001	7-pt Likert scale
Age of Aquisition	18.88	7.79	0.007	20.04	1.77	< 0.001	7-pt Likert scale
Length	142.91	2.95	< 0.001	148.20	2.25	< 0.001	characters

Table 2: Results of 1-stage mixed linear regression (crossed maximal random effects by user and word, participant; $N = 100$) and two-stage regression (participant $N = 462$) analyses.

effects. We further validate these effects with t-tests on the recovered parameter estimates from our two-stage model — **Left**: $t = 8.05$, $p < 0.001$; **Right**: $t = 2.57$, $p = 0.01$; **Left** vs. **Right**: $t = 3.22$, $p = 0.001$. This result is of particular interest for two reasons. Firstly, it demonstrates that in-context predictability plays a significant role even when race prompts are provided prior to the race suggesting prediction is a somewhat automatic part of production and processing. Secondly, the opposite relative effect strength (i.e., **Right** $<$ **Left**) has been demonstrated in spoken language production. While different modalities of language production seem sensitive to overlapping factors they need not show the same fine-grained patterns of sensitivity, and typing seems especially sensitive to **Left** context.

We also show the effects of each of the other predictors in Table 2. Because these latter predictors are treated as linear in our analysis without rigorous verification, these results are somewhat provisional. Several factors bear significantly on typing time. Unsurprisingly, length had the largest effect on typing time. The two predictors with the next highest absolute t-values were orthographic neighborhood density and unigram surprisal. Taken together, these suggest that while humans are well tuned to the statistics of their language and use these statistics to coordinate typing behavior, they are also subject to influence from orthographic factors (neighborhood density) to a large degree. We see significant effects from several other Glasgow norms (Age of Acquisition, Gender Association, Dominance, Semantic size). Interestingly, repetition did not significantly improve predicted typing time, unlike in spoken language production (Kahn & Arnold, 2013).

General Discussion

Our results provide what are, to our knowledge, the first demonstration of contextual probability effects on word-level typing speed. We find that typists are reliably sensitive to both the forward (**Left**) and backward (**Right**) predictability of words in the text prompts provided on Typeracer. Similar to well-documented effects in spoken language production, words that are more probable in context take less time to produce. But in contrast to spoken language production, the effect of predictability based on left context is a stronger determinant of word typing time than predictability based on right context (Bell et al., 2009), and repetition had no effect, which is at odds with the well-attested effect in spoken production (Bell et al., 2009; Lam & Watson, 2010; Kahn & Arnold, 2013). Moreover, we find that other word-level factors studied in spoken language production, or analogous to those studied in spoken language production, predict typing time as well. Thus, efficiency in typed language production turns out to have similar overall sensitivity to the factors that influence spoken language production—but the weighting of these factors is different. Notably, unlike most past work in spoken production, we estimate in-context predictability with LSTM language models that have access to the *entire* (**Left** or **Right**) context (as opposed to bigram estimates as in e.g. (Bell et al., 2009)). It is not clear how this modeling difference would affect the difference in sensitivities demonstrated in typing/spoken language.

There are several limitations of the data presented here, most of which stem from the fact that data collection on `Typeracer.com` was not designed with psycholinguistic analysis in mind. First, with a self-selected population we do not have fine grain control over participant demographics. Furthermore, past work shows that there is a large effect of motor control that bears on typed production that our word-level analysis does not address. Looking ahead, the character-level typing time information available in TypeRacer can provide a testbed for competing hypotheses for the role of mo-

tor control in typing. Morpheme and character-level conditional probability can be analyzed with similar methods to the ones we deploy at the word level and we leave these avenues to future work. Additionally, because the text is displayed on-screen while participants type, typing time delays are not fully disentangled from delays due to reading time (Smith & Levy, 2013). For a full characterisation of typing behavior, targeted in-lab experiments will likely be a useful supplement to the analysis approach demonstrated here. Lastly, because languages can have quite variable typing systems, the interpretations of effects reported here would likely hold most strongly for language with typing systems most similar to English (i.e., one character, one keystroke typing systems). Still, massive, publicly available datasets like the one examined here permit analysis at a scale far beyond what is possible in the lab, helping scientists ask qualitatively different questions than those possible with small-scale experimentally collected data (Griffiths, 2015).

Acknowledgements

TE gratefully acknowledges support from the GEM consortium and the National Science Foundation Graduate Research Fellowship under Grant No. 1745302. RL gratefully acknowledges support from NSF award IIS1815529, a Google Faculty Research Award, a Newton Brain Science Award, and MIT's Quest for Intelligence. We thank anonymous reviewers and the participants of the 34th CUNY conference for helpful feedback and discussions.

References

Arnold, J. E. (2008). Reference production: Production-internal and addressee-oriented processes. *Language & Cognitive Processes*, *23*(4), 495–527.

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

Aylett, M., & Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *Journal of the Acoustical Society of America*, *119*(5), 3048–3058.

Balota, D. A., & Chumbley, J. I. (1985). The locus of word-frequency effects in the pronunciation task: Lexical access and/or production? *Journal of Memory and Language*, *24*(1), 89–106.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013, April). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J. Mem. Lang.*, *68*(3).

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*(1), 92–111.

Boomer, D. S. (1965). Hesitation and grammatical encoding. *Language and Speech*, *8*(3), 148–158.

Cohen Priva, U. (2010). Constructing typing-time corpora: A new way to answer old questions. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 32).

Eisape, T., Zaslavsky, N., & Levy, R. (2020). Cloze Distillation: Improving Neural Language Models with Human Next-Word Predictions. In *Proceedings of the 24th conference on computational natural language learning* (pp. 609–619). Association for Computational Linguistics. doi: 10.18653/v1/2020.conll-1.49

Gahl, S., Yao, Y., & Johnson, K. (2012). Why reduce? phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, *66*(4), 789–806.

Galati, A., & Brennan, S. E. (2010). Attenuating information in spoken communication: For the speaker, or for the addressee? *Journal of Memory and Language*, *62*(1), 35–51.

Gelman, A. (2005). Two-stage regression and multilevel modeling: A commentary. *Political Analysis*, *13*(4), 459–461. doi: 10.1093/pan/mpi032

Gentner, D. R. (1982). Evidence against a central control model of timing in typing. *Journal of Experimental Psychology: Human Perception & Performance*, 793–810.

Gentner, D. R., Larochelle, S., & Grudin, J. (1988). Lexical, sublexical, and peripheral effects in skilled typewriting. *Cognitive Psychology*, *20*(4), 524–548.

Griffiths, T. L. (2015, February). Manifesto for a new (computational) cognitive revolution. *Cognition*, *135*, 21–23.

Hall, K. C., Hume, E., Jaeger, T. F., & Wedel, A. (2018). The role of predictability in shaping phonological patterns. *Linguistics Vanguard*, *4*(s2).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In J. Bybee & P. Hopper (Eds.), *Probabilistic relations between words: Evidence from reduction in lexical production* (pp. 229–254). Benjamins.

Kahn, J. M., & Arnold, J. E. (2013, October). Articulatory and lexical repetition effects on durational reduction: Speaker experience vs. common ground. *Language Cognition and Neuroscience*, *30*(1-2), 103–119.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., ... Herbst, E. (2007, June). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions* (pp. 177–180). Prague, Czech Republic: Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/P07-2045`

Lam, T. Q., & Watson, D. G. (2010, December). Repetition is

easy: why repeated referents have reduced prominence. *Mem. Cognit.*, *38*(8), 1137–1146.

Levelt, W. J. M. (1993). *Speaking: From intention to articulation*. MIT Press.

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In *Proceedings of the 20th conference on Neural Information Processing Systems.*

Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology*, *88*, 22–60.

Marian, V., Bartolotti, J., Chabal, S., & Shook, A. (2012). Clearpond: Cross-linguistic easy-access resource for phonological and orthographic neighborhood densities. *PloS one*, *7*(8), e43230.

Meinhardt, E., Baković, E., & Bergen, L. (2020, July). Speakers enhance contextually confusable words. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 1991–2002). Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.acl-main.180 doi: 10.18653/v1/2020.acl-main.180

Merity, S., Xiong, C., Bradbury, J., & Socher, R. (2016). Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Momma, S., Slevc, L. R., & Phillips, C. (2016). The timing of verb selection in japanese sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(5), 813.

Oldfield, R., & Wingfield, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology*, *17*(4), 273–281.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc. Retrieved from http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2005). Lexical frequency and acoustic reduction in spoken dutch. *The Journal of the Acoustical Society of America*, *118*(4), 2561–2569.

Roser, M., & Ortiz-Ospina, E. (2016). *Literacy.* Retrieved from https://ourworldindata.org/literacy

Scarborough, R. (2012). Lexical similarity and speech production: Neighborhoods for nonwords. *Lingua*, *122*(2), 164–176.

Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., & Sereno, S. C. (2019). The glasgow norms: Ratings of 5,500 words on nine scales. *Behavior research methods*, *51*(3), 1258–1270.

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, *133*(1), 140–155.

Shriberg, E. (2001). To 'errrr' is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 153–169.

Shriberg, E., & Stolcke, A. (1996). Word predictability after hesitations: A corpus-based study. In *Proceedings of the international conference on spoken language processing, philadelphia.*

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

Speer, R., Chin, J., Lin, A., Jewett, S., & Nathan, L. (2018, October). *Luminosoinsight/wordfreq: v2.2.* Retrieved from https://doi.org/10.5281/zenodo.1443582 doi: 10.5281/zenodo.1443582

Terzuolo, C. A., & Viviani, P. (1980). Determinants and characteristics of motor patterns used for typing. *Neuroscience*, *5*(6), 1085–1103.

Umeda, N. (1977). Consonant duration in American English. *The Journal of the Acoustical Society of America*, *61*(3), 846–858.

Vitevitch, M. S., Armbrüster, J., & Chu, S. (2004). Sublexical and lexical representations in speech production: effects of phonotactic probability and onset density. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 514.

Watson, D. G., Arnold, J. E., & Tanenhaus, M. K. (2008). Tic tac toe: Effects of predictability and importance on acoustic prominence in language production. *Cognition*, *106*(3), 1548–1557.

Weingarten, R., Nottbusch, G., & Will, U. (2004). Morphemes, syllables, and graphemes in written word production. In T. Pechmann & C. Habel (Eds.), *Trends in linguistics studies and monographs* (Vol. 157, pp. 529–572). Berlin: Mouton de Gruyter.

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. P. (2020). On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42nd annual meeting of the Cognitive Science Society* (pp. 1707–1713). Retrieved from https://cognitivesciencesociety.org/cogsci20/papers/0375/0375.pdf

Wood, S. (2006). *Generalized additive models: an introduction with R*. CRC press.