INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Time-to-Event Models for Analyzing Reaction Time Sequences

Louis ten Bosch , Lou Boves

Centre for Language Studies, Radboud University, Nijmegen, The Netherlands
l.tenBosch@let.ru.nl, l.boves@ru.nl

Abstract

We investigate reaction time (RT) sequences obtained from lexical decision experiments by applying Time-to-Event modelling (Survival Analysis). This is a branch of statistics for analyzing the expected duration until one or more events happen, associated with a set of potential 'causes' (in our case the decision for a 'word' judgment as a function of conventional predictors such as lexical frequency, stimulus duration, reduction, etc.). In this analysis, RTs are considered a by-product of an (unobservable) cumulative incidence function that results in a decision when it exceeds a certain threshold.

We show that Survival Analysis can be effectively used to narrow the gap between data-oriented models and process-oriented models for RT data from lexical decision experiments. Results of this analysis technique are presented for two different RT data sets. The analysis reveals time-varying patterns of predictors that reflect the differences in cognitive processes during the presentation of auditory stimuli.

Index Terms: multi-state models, reaction times, psycholinguistics, auditory lexical decision

1. Introduction

Reaction Time (RT) is among the most frequently used behavioral measures in experiments that aim to understand how humans react to stimuli of all sorts, and specifically to linguistic stimuli. In this paper we focus on lexical decision experiments, and auditory lexical decision in particular. A lexical decision ('is this a word or not?') can be regarded a central functionality of human speech perception. The final goal of those experiments is to uncover the contribution of a multitude of features of the stimuli to the reaction times (RTs), as a 'measure of difficulty' of this task. Clever manipulation of those features should make it possible to draw conclusions about cognitive processes that drive the reactions. For over a decade, RTs obtained in lexical decision experiments have been modeled using linear mixed-effect models (LMM) (e.g., [1, 2, 3, 4]). The most important advantage of LMMs over traditional linear models in many fields of science is well known (e.g., [5]): the random structure in mixed-effects models makes it possible to estimate more conservative statistical models and to consider e.g. items and participants as related to normally distributed perturbations on top of the predictions from the fixed structure.

However, LMMs predict the values of a dependent variable on the basis of a (potentially complex) set of predictor variables without intending to uncover the –also potentially complex– time course of the processes that generated the dependent variable data in the first place. In other words: LMMs are data models, not process models [6]. One implication is that LMM models are not able to account for the fact that the role of some predictors may change over time. For example, [7] showed that the role of lexical frequency may change during the course of the processing of individual stimuli. Generalized Additive Mixed Models (GAMM) solve the problem of 'static predictors' by replacing fixed predictors with spline functions in modeling data consisting of time series (e.g., [8] for air pollution data and [9] for EEG traces). Despite the fact that RTs recorded from individual participants in lexical decision experiments are indeed time series, and GAMMs could be used for removing local speed effects from those time series, it is prohibitively difficult to use GAMMs for uncovering the time-dependent effect of predictors on the mean RTs of a number of participants for a collection of stimuli.

It is generally agreed that the performance of participants in an auditory lexical decision experiment involves several cognitive processes. Exactly which processes are assumed to be involved depends on the theory that one espouses, but most 'schools' will accept that at least three different processes are involved: (1) phonetic form decoding, (2) lexical-semantic access and (3) decision making. The temporal relation between those processes is also subject of discussion. In the most recent version of a computational model of human speech processing DIANA ([10, 11]) it is assumed that at least the processes 'Activation' (similar to phonetic form decoding) and 'Decision' (similar to lexical-semantic access) and perhaps also 'Execution' (pressing a button) can proceed in parallel, at least for part of the time between stimulus onset and button press. Under this model, every participant can be in three different states while processing a given stimulus. Modeling the sequences in which these states are visited and the times at which a transfer from one state to the next takes place surely is more informative than just modeling the observed RTs.

Cognitive processes can be studied by involving other techniques, such as the use of neuro-physiological measurements. There is a long tradition and a vast literature on using brain imaging and electrophysiological techniques to elucidate cognitive processes and their timing in psycholinguistics (e.g., [12, 13]). However, these techniques come with problems and limitations of their own (e.g., see the discussion about possible interpretations of the N400 effect [14]). Therefore, it seems relevant to look for statistical approaches that may be closer to process models and that can be applied to existing data sets of behavioral experiments. One potentially interesting family of techniques that is widely used in economics and medical (especially epidemiological) research is known as 'time-to-event' or 'survival analysis' (e.g. [15, 16].

Survival Analysis has been used in psycholinguistic and psychological research before. For example, [17] and [18] used the survival approach for analyzing fixation times in eye tracking studies. In [19] the approach was used in an analysis of language acquisition. Perhaps most relevant in the context of lexical decision, [20] used Survival Analysis to analyze and understand the accumulation of information underlying reaction times in psychological tests.

In this paper we investigate whether Survival Analysis can be used to narrow the gap between data models and process

http://dx.doi.org/10.21437/Interspeech.2021-1408

models for RT data from two lexical decision experiments using auditory stimuli in isolation. In doing so (and in contrast with [20]), we will focus on techniques that allow investigating how the impact of stimulus features might change during the processing of a stimulus [21].

2. Survival Analysis

Survival Analysis (SurvA) is a branch of statistics for predicting the moment in time when specific events are expected to occur. The techniques are also known as 'reliability analysis' in engineering, 'duration modelling' in economics, and 'event history analysis' in sociology [22]. The techniques aim to answer questions such as 'what is the proportion of stimuli with $RT > T_0$?' (for some T_0) and 'are there particular features of items that increase or decrease the probability of observing an $RT > T_{\text{threshold}}$?'. The reader is kindly referred to [15] for an introduction to a wide range of different SurvA techniques. In this paper we focus on techniques that allow for modeling RT data when there are multiple processes that contribute to the results and that can operate in parallel or in sequence.

2.1. Multi-state Models

Models of spoken word comprehension can differ in the number of (cognitive) states that are hypothesized. For example, a theory might consider lexical access and semantic access as two separate processes. Row A and B in Figure 1 show different ways in which a three-state model such as DIANA can be visualized. The top row (A) conforms with the assumption that Activation (phonetic decoding) and Decision (lexical-semantic access) may operate at least partly in parallel. It allows for the possibility to go from Activation to Execution directly (as might be the case when Activation yields a single hypothesis). The middle row (B) shows a model in which Activation and Decision operate in sequence, and Execution is split in two states ('word' or 'pseudoword (non-word)'). The model shown in the bottom row (C) takes a very different approach. Its serial architecture implies that states operate essentially in sequence, but the number of states is unknown and the actual operations performed in the states are open to interpretation. In this paper we focus on the architecture in the bottom row. In other words, we start out without strong assumptions about the actual architecture of the cognitive process. Instead, we try to infer the operations in the states from analyses of RT data from two fairly large-scale lexical decision experiments with different sets of predictors.

2.1.1. BALDEY

BALDEY [23] is an experiment in which twenty native Dutch listeners (10 male, 10 female, 18 to 23 years) without reported hearing problems were paid to make lexical decisions. For each of the 20 participants, the experiment consisted of 10 sessions, one per week. Each participant made lexicality decisions on a total of 5541 stimuli, about half of which were pseudo words. In total, BALDEY contains over 110,000 reaction times and judgments. The real words form a rich set of morphological constructions, including Noun-Noun, Noun-Adjective and Adjective-Noun compounds. The pseudowords were constructed using the same 'morphological' patterns.

In the analyses in this paper we used the log-transformed RT sequences recorded in the individual BALDEY sessions. In the BALDEY data we are specifically interested in effects of the lexical frequency of the existing words, as well as the effect of duration and the morphological make-up of the words.

2.1.2. Cognates experiment

The 'Cognates' experiment [24] involved lexical decisions by 31 advanced Dutch learners of English and 38 native English listeners. The stimuli were trisyllabic mono-morphemic words with lexical stress on the first syllable. Half of the 92 target stimuli were cognates for the Dutch learners. In addition, all 92 target stimuli were pronounced (by a male native speaker of English) in two forms: once with a canonical pronunciation and once with a (heavily) reduced medial syllable.

In the analysis of the data of the cognates experiment we are interested in the effects of lexical frequency, reduction and the cognate status of the words, as well as potential differences between native and non-native listeners.

2.1.3. Defining the 'event'

The use of time-to-event models requires the definition of an observable 'event'. In lexical decision experiments the concept of event can be defined in several ways. One definition may be whether a decision was correct (our use); alternatively, 'event' might be defined as a lexicality judgment for a stimulus. The definition of 'event' has implications for the status of predictor variables. For example, lexical frequency can be used as a predictor for the subset of word stimuli, but it has no useful value for pseudoword stimuli.

2.1.4. The implementation of the survival models

There is a large number of survival models with different features. In this study we use several functions from the Python package `lifelines` [21], specifically the function `PiecewiseExponentialRegressionFitter` which makes it possible to divide the time axis into an arbitrary (user-defined) number of subsequent intervals of arbitrary length, for each of which a parametric estimate of the survival function is computed. The function performs a regression which makes it possible to estimate the combined effect of multiple predictors per interval. From a mathematical point of view, survival regression models are related to generalized additive mixed models [25] but based on the concept of multi-state accumulative incidence.

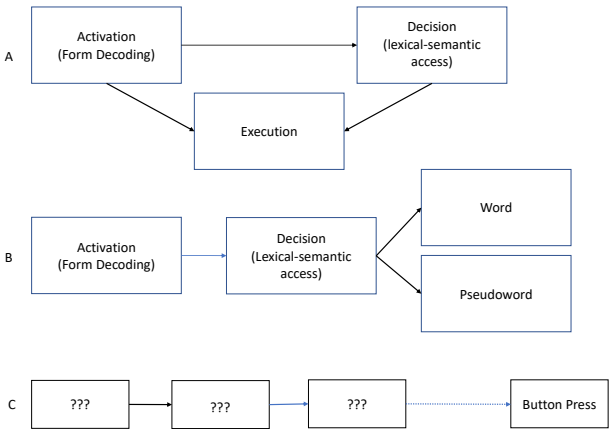

Figure 1: *Schematic diagram of various state transitions in a model of auditory lexical decision. In all cases, the input is the unfolding audio signal (left out from the diagram).*

3. Results

3.1. BALDEY

Using `PiecewiseExponentialRegressionFitter` we estimated survival models for the reaction times of the words and pseudowords in BALDEY. Separate models were fitted for the words and pseudowords, because the predictor *word frequency* does not exist for the pseudowords. The *event* in estimating the models is defined as the correctness of the responses. Separate pairs of models were fitted for log(RT_{onset}) and log(RT_{offset}), in which the reaction time is measured from stimulus onset or stimulus offset, respectively. The predictors of interest are `log(worddur)`, `log(frequency)` and `compound type`. The `log(RT)` range was divided into 20 equal-length intervals spanning the range between 0 and 8 (the maximum log(RT_{onset})). The same time axis segmentation was used for the time interval between stimulus onset and stimulus offset.

Figure 2 shows the effect of the stimulus duration on the expected RT. The blue (for words) and cyan (for pseudowords) traces show the effect on RT_{onset}, while the red (words) and magenta (pseudowords) traces show the effect on RT_{offset}. The traces represent the contribution of the predictor `log(stimulus_duration)` similar to the βs in linear (mixed) models, over and above a baseline that accounts for the summed effects of all predictors similar to the intercept in LMMs. The two lines corresponding to RT_{onset} lie above the x-axis ($y > 0$), showing that log(RT_{onset}) increases with `log(stimulus_duration)`. The opposite is true for log(RT_{offset}), which decreases with `log(stimulus_duration)`. In numerical terms the contribution to log(RT_{onset}) is a bit larger. The time axes for onset and offset are different, because the time between stimulus offset and the fastest RT is much shorter that the time between stimulus onset and the fastest RT. The traces in Figure 2 and all subsequent figures contain two components. The average value corresponds to the 'static' effect of a predictor on the expected RT. What we are interested in here are the variations above and below that average value as a function of predicted RT.

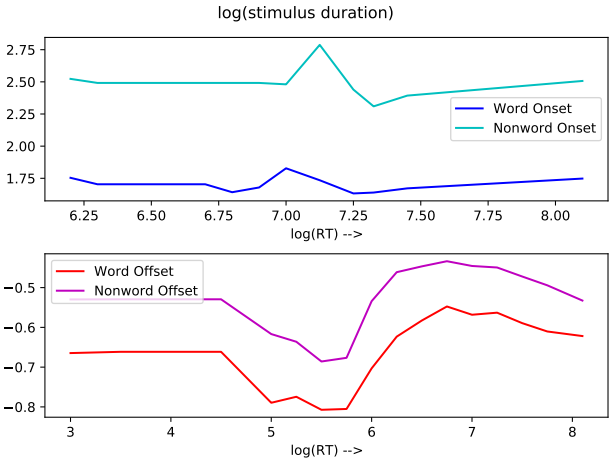

Figure 2: *The effect of the duration of the stimuli on expected reaction time when event is correctness of the response.*

For log(RT_{onset}) (upper panel) the overall effect of stimulus duration is much larger for the pseudowords than for words. In addition, it can be seen that fluctuations over time occur earlier in the words than in the pseudowords, suggesting that cognitive processes related to stimulus duration take effect earlier in the words than in the pseudowords. This holds both for effects that increase and decrease the contribution of stimulus duration to expected RT.

For log(RT_{offset}), the time-varying effect of the stimulus duration is opposite: larger for the words than for the pseudowords. Short RTs measured from stimulus offset appear to be extra short ($\log(RT) \approx 5$ corresponds to a linear RT of $\approx$ 150 ms). Longer RTs ($\log(RT) \approx 7$, that is, $RT \approx 1100$ ms) lead to an extra increase of expected RT, while for very long RTs the effect is back to its baseline value. Words and pseudowords seem better synchronized, compared to RT_{onset}.

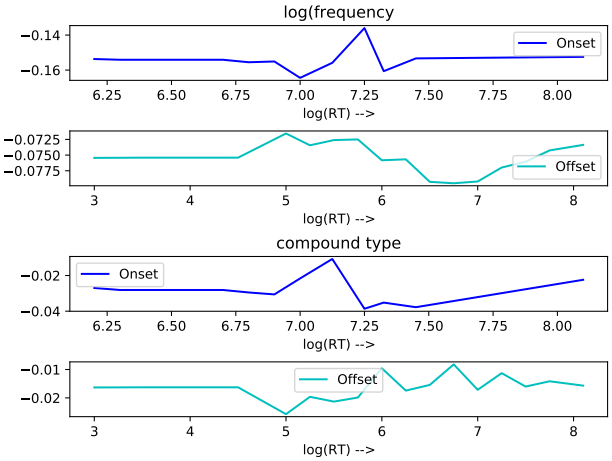

Figure 3: *The effect of log(frequency) and compound type of the stimuli on expected reaction time when event is correctness of the response.*

Figure 3 shows similar data for `frequency` (top panel) and `compound type` (bottom panel), two predictors that are only available for the word stimuli. Overall, `frequency` subtracts a small (but statistically significant) amount from the baseline prediction of log(RT), more so for log(RT_{onset}). In addition, `frequency` tends to increase the shortest log(RT_{offset}) (on a linear scale about 150 ms), to decrease log(RT_{offset}) at around 800 ms, and then again to increase for very large RTs, $RT_{offset} \approx 2500$ ms. For log(RT_{onset}) we only see a shortening effect for $RT_{onset} \approx 1000$ ms and a lengthening effect for RTs around 1600 ms. All traces shown here for BALDEY show significant effects except for `compound type`, both for RT_{onset} and RT_{offset}, which are shown for the sake of completeness.

3.2. The cognates experiment

In the *cognates* experiment we are interested in several effects: (a) natives versus learners, (b) lexical frequency, (c) stimulus duration, (d) reduction and (e) cognate status. Because the function `PiecewiseExponentialRegressionFitter` does not offer a strata option (which would allow to group the natives and learners in two strata) we decided to build separate models for the natives and learners, and compare the results. As with the BALDEY data we analyzed both RT_{onset} and RT_{offset} for the correctly judged target stimuli. The results for stimulus duration and lexical frequency are shown in Figure 4.

From Figure 4 it can be seen that stimulus duration has a similar effect as in the BALDEY data: In the RT_{onset} data longer stimulus durations -predictably- yield longer RTs, while the inverse is true for RT_{offset}. The trajectories for the natives and learners are remarkably similar, except for the overall size of the

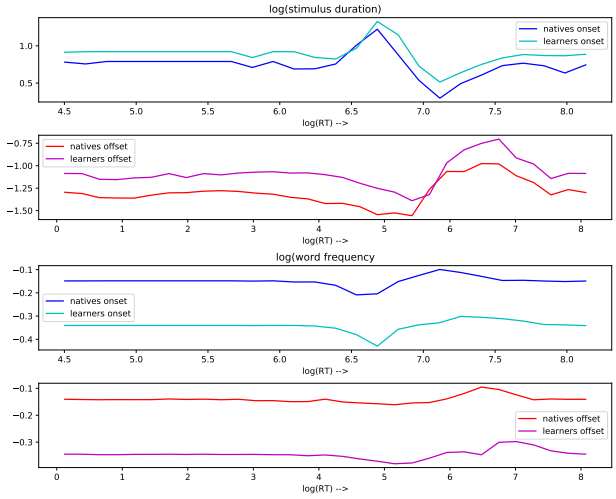

Figure 4: *The time dependence of the predictors* `stimulus duration` *(top) and* `logFrequency` *(bottom) on* RT_{onset} *and* RT_{offset}.

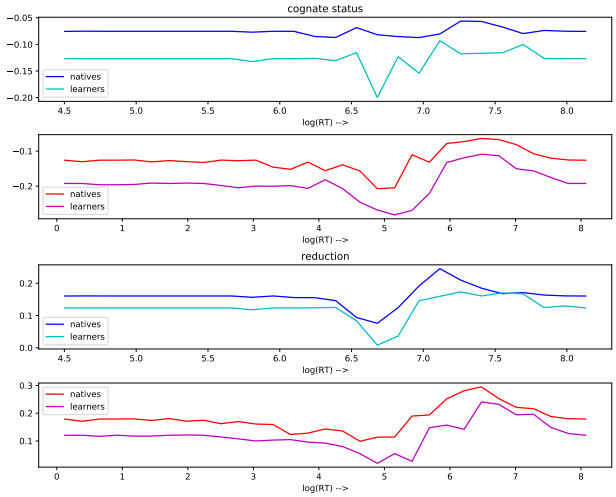

Figure 5: *Time dependence of the factors* `cognate` *(top) and* `reduction` *on* RT_{onset} *and* RT_{offset} *in the cognates data.*

effects. For linear RTs_{onset} of around 750 ms we see an increase of the effect of stimulus duration, while the effect is smaller for $RT_{onset} \approx 1200$ ms. In the data for RT_{offset} the overall effect of stimulus duration is substantially larger (more negative, corresponding with shortening) for the natives. In both groups there is an extra shortening for $RT_{offset} \approx 250$ ms and a lengthening effect for $RT_{offset} \approx 800$ ms.

While the overall effect of logFrequency is highly significant, both for RT_{onset} and RT_{offset}, there is little variation of the effect as a function of log(RT) proper, except perhaps for $RT_{onset} \approx 800$ ms for the learners. Overall, the effect of logFreq is substantially larger for the learners. Arguably, this can be attributed to the fairly small range of this predictor ($1.86 <$ logFreq < 5.61) that would affect learners more than natives.

Figure 5 shows the effects of the factors `cognate` (top) and `reduction` (bottom). For the natives the effect of `cognate` on RT_{onset} is very small, and there is hardly a change as a function of the length of RT. For the learners there appears to be a relative shortening for $RT_{onset} \approx 800$ ms. Surprisingly, for RT_{offset} the effect of cognate status appears to change as a function of time also for the natives, for whom the concept of *cognate* is nonexistent. The effect is smaller, but almost perfectly synchronized with the effect for the learners. This suggests that there is more to the cognate stimuli than can be explained by the fact that these words happen to be Dutch-English cognates. That effect is not lexical frequency: the distribution of the lexical frequency of the cognate stimuli does not differ significantly from the distribution of the other word stimuli.

The change as a function of predicted RT of the factor `reduction` is very similar for the two groups. Somewhat surprisingly, this effect is larger for the natives. The extra lengthening due to reduction is smaller for short RTs and larger for longer RTs. That effect holds for both onset and offset.

We also looked at the effect of the factor `word/nonword` (not shown in a figure). While the overall effect of this factor was highly significant, its change over time was very small. As expected, the effect of this factor was much larger (corresponding to shorter RTs for onset and offset) for the natives than for the learners.

4. Discussion and future research

Our results show that Survival Analysis can uncover subtle but significant changes over time in the regression coefficients (β's) for a number of relevant predictors for RTs coupled to some 'event' (here: the correctness of a decision) in human auditory lexical decision data. For many predictors such as word duration and word frequency but also cognate status, β's vary significantly during the unfolding of the stimulus. Other predictors, such as word class and compound type did not show such variation, which may imply that such meta-level features do not affect linguistically naive listeners' processing. In a principled way, survival model are able to relate an unobservable accumulative incidence function (interpreted as an internal confidence growth function, to which the overt behavioral RT is related as a side effect) with time-varying regression. SurvA provides a quantitative way to show that stimulus features that are used as proxies for the cognitive processes playing a role in word processing change over time.

This study is limited to one member of the family of SurvA techniques that allows breaking up the time axis into arbitrary segments. Arguably, that is not an ideal way for analyzing processes that may in part operate in parallel. We leave attempts to link segments in which specific predictors vary to (combinations of) specific processes to future research, based on a larger number of data sets from experiments with similar goals. Also, we plan to investigate the use of other members of the SurvA family that are able to estimate the parameters of the more complex architectures in Figure 1, and to study the relation between GAMMs and SurvAs for the modeling and interpretation of RTs in psycholinguistic experiments in more depth.

A potential limitation of SurvA for the study of the way in which the effect of stimulus features changes over time is that these techniques are unable to shed light on processes that occur before the shortest RT, because up to that time all stimuli are 'alive'. Still, we expect that SurvA can shed light on links between RTs and the cognitive processes that operate after stimulus offset. This will help to improve the Decision component in DIANA[10], one of the few computational models of spoken word comprehension that processes real audio. For an in-depth analysis of the processes before stimulus offset we will take recourse to electrophysiological (e.g., [26]) and brain imaging techniques.

SurvA scripts are available upon request from the authors.

5. References

[1] S. Lo and S. Andrews, "To transform or not to transform: Using generalized linear mixed models to analyse reaction time data," *Frontiers in Psychology*, vol. 6, 2015.

[2] H. Baayen and P. Milin, "Analyzing Reaction Times," *International Journal of Psychological Research*, vol. 3, no. 2, pp. 12 – 28, 2010.

[3] L. ten Bosch, L. Boves, and K. Mulder, "Analyzing reaction time and error sequences in lexical decision experiments," in *Proceedings Interspeech*, Graz, Austria, 2019, pp. 2280 – 2284.

[4] L. ten Bosch, M. Ernestus, and L. Boves, "Analyzing reaction time sequences from human participants in auditory experiments," in *Proc. Interspeech 2018*, 2018, pp. 971–975.

[5] X. A. Harrison, L. Donaldson, M. E. Correa-Cano, J. Evans, D. N. Fisher, C. Goodwin, B. S. Robinson, D. J. Hodgson, and R. Inger, "A brief introduction to mixed effects modelling and multi-model inference in ecology," *PeerJ*, vol. 6, p. e4794, 2018.

[6] A. Lugmayr, B. Stockleben, C. Scheib, and M. Mailaparampil, "Cognitive big data - survey and review on big data research and it's implications: What is really new in big data?" *Journal of Knowledge Management*, vol. 21, 02 2017.

[7] D. Dahan, J. S. Magnuson, and M. K. Tanenhaus, "Time course of frequency effects in spoken-word recognition: Evidence from eye movements," *Cognitive Psychology*, vol. 42, no. 4, pp. 317 – 367, 2001.

[8] Y. Chuang, S. Mazumdar, T. Park, G. Tang, V. C. Arena, and M. J. Nicolich, "Generalized linear mixed models in time series studies of air pollution," *Atmospheric Pollution Research*, vol. 2, no. 4, pp. 428–435, 2011.

[9] K. Mulder, L. ten Bosch, and L. Boves, "Analyzing EEG signals in auditory speech comprehension using temporal response functions and generalized additive models," in *Proc. Interspeech 2018*, 2018, pp. 1452–1456.

[10] L. ten Bosch, L. Boves, and M. Ernestus, "DIANA: towards computational modeling reaction times in lexical decision in north american english," in *Proceedings of Interspeech*, Dresden, 2015, pp. 1576 – 1580.

[11] F. Nenadić, L. ten Bosch, and B. V. Tucker, "Implementing DIANA to model isolated auditory word recognition in english," in *Proc. Interspeech 2018*, 2018, pp. 3772–3776.

[12] J. Rommers and K. D. Federmeier, "Electrophysiological methods," in *Research Methods in Psycholinguistics and the Neurobiology of Language: A Practical Guide*, A. M. B. de Groot and P. Hagoort, Eds. Hoboken, NJ: Wiley, 2017, pp. 247–265.

[13] J. M. Rodd and M. H. Davis, "How to study spoken language understanding: a survey of neuroscientific methods," *Language, Cognition and Neuroscience*, vol. 32, no. 7, pp. 805–817, 2017.

[14] M. Kutas and K. D. Federmeier, "Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP)," *Annual Review of Psychology*, vol. 62, pp. 621–647, 2011.

[15] P. Schober and T. R. Vetter, "Survival analysis and interpretation of time-to-event data: The tortoise and the hare," *Anesthesia & Analgesia*, vol. 127, no. 3, pp. 792 – 798, 2018.

[16] L. Meira-Machado, J. de Uña Alvarez, C. Cadarso-Suárez, and P. K. Andersen, "Multi-state models for the analysis of time-to-event data," *Statistical methods in medical research*, vol. 8, no. 2, p. 195–222, 2009.

[17] M. Leinenger, "Survival analyses reveal how early phonological processing affects eye movements during reading," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 45, no. 7, 2019.

[18] M. Nilsson and J. Nivre, "A survival analysis of fixation times in reading," in *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, ser. CMCL '11. USA: Association for Computational Linguistics, 2011, p. 107–115.

[19] M. Ota and S. J. Green, "Input frequency and lexical variability in phonological development: a survival analysis of word-initial cluster production," *Journal of Child Language*, vol. 40, no. 3, p. 539–566, 2013.

[20] J. Ranger and J.-T. Kuhn, "Modeling information accumulation in psychological tests using item response times," *Journal of Educational and Behavioral Statistics*, vol. 40, no. 3, pp. 274–306, 2015.

[21] C. Davidson-Pilon et al. (2020) Camdavidsonpilon/ lifelines: 0.25.10. [Online]. Available: https://github.com/ CamDavidson-Pilon/lifelines/

[22] "Survival analysis," https://en.wikipedia.org/wiki/Survival_analysis, accessed 15 March 2021.

[23] M. Ernestus and A. Cutler, "BALDEY: A database of auditory lexical decisions," *Quarterly Journal of Experimental Psychology*, vol. Advance online publication, 2015.

[24] K. Mulder, G. Brekelmans, and M. Ernestus, "The processing of schwa reduced cognates and noncognates in non-native listeners of English," in *Proceedings of the 18th International Congress of Phonetic Sciences [ICPhS 2015]*, 2015, pp. 1 – 5.

[25] A. Bender, A. Groll, and F. Scheipl, "A generalized additive model approach to time-to-event analysis," *Statistical Modelling*, vol. 18, no. 3-4, pp. 299–321, 2018.

[26] A. Burgess, "Towards a unified understanding of event-related changes in the EEG: The firefly model of synchronization through cross-frequency phase modulation," *PLoS ONE*, vol. 7, 2012. [Online]. Available: https://doi.org/10.1371/journal.pone.0045630