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

Models of reaction times in auditory lexical decision: RTonset versus RToffset

Sophie Brand[1], *Kimberley Mulder*[2], *Louis ten Bosch*[3], *Lou Boves*[3]

[1]Hotel Management School Maastricht, the Netherlands
[2]Utrecht University, the Netherlands
[3]Centre for Language Studies, Radboud University, Nijmegen, the Netherlands
swmbrand@gmail.com, k.mulder@uu.nl,l.tenbosch@let.ru.nl, l.boves@ru.nl

Abstract

We investigate how the role of predictors in models of reaction times in auditory lexical decision experiments depends on the operational definition of RT: whether the time is measured from stimulus onset or from stimulus offset. In a large body of literature, RTs are measured from the onset of the stimulus to the start of the response (often a button press or an oral response). The rationale behind this choice is that information about the stimulus becomes available to the listener starting at onset. Alternatively, the RT from offset is less dependent on stimulus duration and is assumed to focus on those cognitive processes that play a role late(r) in the word and after word offset, when all information is available.

The paper presents RT-onset and RT-offset-based linear mixed effects models for three different lexical decision-based data sets and explains the significant differences between these models, showing to what extent both definitions of reaction time reveal different roles for predictors and how early and later contributions to the overall RT can be differentiated.

Index Terms: reaction times, psycholinguistics, auditory lexical decision, onset, offset

1. Introduction

Lexical decision experiments are a frequently used and classical method in psycholinguistics. A listener's reaction time is assumed to be both related to the complexity of the stimulus and of the task.

This paper discusses how the role of predictors in models of reaction times (RT) in auditory lexical decision experiments depends on the operational definition of RT: whether the RT is measured from stimulus onset (RT_{onset}) or stimulus offset (RT_{offset}). More often than not RTs are measured from the onset of the stimulus to the start of the response (often a button press or an oral response). The rationale behind this choice is that already from the start of the stimulus, all kinds of information about the stimulus become available to the listener. Therefore, it seems plausible, from a processing point-of-view, to measure RT from word onset.

However, one could object that RTs measured from onset depend to a large extent on the length of the stimulus itself (i.e., the word duration). The word duration is known to be a highly significant predictor with positive β in the statistical regression models of RTs measured from stimulus onset (e.g. [1, 2, 3]). The alternative is to measure RTs from the *offset* of the stimulus to the start of the overt response (e.g., [4, 5]. Besides accounting for the length of the stimuli, it might be argued that this method more accurately reflects the specific cognitive processes that take place between word offset and response. The question arises in what respect RTs measured from word onset capture other processes than those measured from word offset. One could argue that at the offset of a word, cognitive processes related to acoustic form of the stimulus have finished and that the time between stimulus offset and response reflects the 'later' or 'higher-level' lexical-semantic or decision processes. This issue is relevant since in the literature [6, 7] the closely related question is raised whether the impact of the lexical frequency of the stimuli is different before and after stimulus offset.

The differences between RT_{onset} and RT_{offset} may also bring to light other types of differences, e.g., between the auditory processing of native and non-native listeners. As non-native listeners generally have more difficulty (due to various reasons) in the processing of auditory stimuli than native listeners, certain cognitive processes may be delayed until after word offset. It might then be that the effects of certain predictors are more pronounced in RT_{offset} models than in RT_{onset} models. In addition, if non-native listeners need more input and/or time to process a certain auditory stimulus, this leaves more room for activation of the native language. This activation could play out in different ways, depending on the stimulus.

When measured from stimulus offset, RTs may be negative since listeners need not hear the complete stimulus to make a response. This reduces the type of transformations that can be applied to RT_{offset}-measurements, such as the conventionally used $\log()$-transformation. However, it appears that decisions before stimulus offset are quite rare. Therefore, in most cases it is acceptable to simply drop these rare cases with early decisions, accompanied by model criticism.

In this paper we investigate how the role of predictors in statistical models of RTs in auditory lexical decision experiments depends on the use of RT_{onset} or RT_{offset}.

2. Method

2.1. Data sets and theory

In this paper we use the data of three lexical decision experiments, viz. BALDEY [1] in which native Dutch listeners judged a large number of Dutch word/pseudoword stimuli, and two experiments that compare the judgments of advanced Dutch learners of English [8] and French [9] with native listeners. Details about the data sets are provided below.

Lexical decision is likely to involve several cognitive processes that can operate in parallel or in series. The top part of Figure 1 shows a hypothetical schematic view of the processes; the bottom part of Figure 1 shows the distributions of RT_{onset} and RT_{offset} of the correctly judged stimuli in the three experiments under analysis. The solid lines are for RT_{onset}, the dashed lines for RT_{offset}. The different shapes of the distributions are a direct result of the design of lexical decision experiments: valid RT_{onset} judgements can only be made after the end of the stimuli. As a consequence, the minimum value of valid RT_{onset} must be larger than the duration of the stimuli.

http://dx.doi.org/10.21437/Interspeech.2021-1700

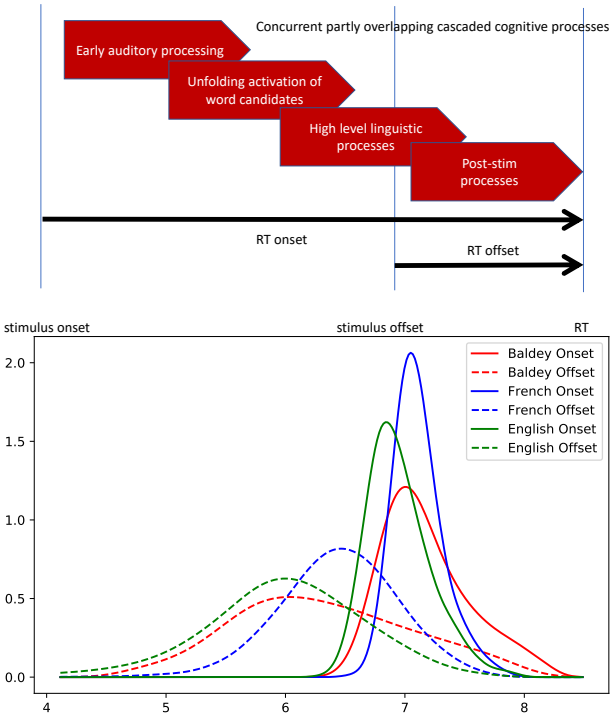

Figure 1: *Multiple processes operate in lexical decision (top). The distributions of* RT_{onset} *are much narrower than those for* RT_{offset} *(bottom).*

2.2. Regression

To investigate whether modelling RT_{onset} and RT_{offset} brings to light interesting differences between early and late effects of predictors, we need statistical methods that can uncover such differences. In this study we use linear mixed effects models with RT_{onset} and RT_{offset} as dependent variable. The direct comparison between models estimated on different dependent variables is not straightforward – measures such as AIC [10] and BIC [11] assume the dependent variable to be the same across models. In addition, the relation between regression on the one hand and a cognitive processing account on the other hand is difficult. In the approach used here, we can study the effects on predictors on different representations of the reaction time.

In our modelling, all `lmer()` models are the result of an iterative reduction of the complexity of the regression, by starting a theoretically defendable regression model with all predictors of interest and control predictors in the fixed structure, and with predictors of interest as random slopes. Next we constrained correlation parameters to zero, and dropped non-significant variance components and their associated correlation parameters from the model, such that the model is actually better supported by information in the data. During this procedure, a failure to converge was taken as a warning that the model was too complex to be properly supported by the data. In addition we avoided overparameterization to avoid the risk of uninterpretable models. However, in order to facilitate the comparison of the two experiments with Dutch learners of English and French, we decided to settle for a single model that provided a near-optimal fit for both data sets.

Comparing models of RT_{onset} and RT_{offset}

Instead of estimating separate `lmer` models for RT_{onset} and RT_{offset}, we estimate one `lmer` model on the combined set of RT onset and offset measurements, and include a predictor `offset` with two factor levels, similar to the set-up of a classical ANOVA between-condition design (see, e.g., [12]). This set-up needs care since it may lead to type I errors: in this case the measurement in one condition (onset) is dependent on the measurement in the other condition (offset) via the stimulus duration. In line with recommendations in [13], we report statistical results without grouping these based on arbitrary p-value thresholds; instead, t-values are presented as continuous quantities ([13], section 2).

Removing local speed effects

When modeling RTs, it is necessary to take into account the 'local speed effect', according to which a substantial part of the observed variation in the RT sequence is due to local trends [14]. To remove those effects it is necessary to regard the RTs as a sequence. In doing this, we considered only sequences within separate blocks in a session. First, we replaced $RT < 250$ ms by the minimum RT and all non-response values by the median RT of the preceding RTs in the session. Next, we removed the trend in the sequence by subtracting a fifth-order Chebyshev fit to the sequence, followed by a robust median filtering. Because trend removal shifts the average RT value in a session to zero, we shifted the detrended and filtered RTs such that the overall average becomes equal to the raw average (with outliers and non-responses removed).

3. Analyses

3.1. An experiment with Dutch learners of English

The first set of data comes from [8]. Their stimuli consisted of 196 real mono-morphemic English words and 200 pseudowords. Of the real English words, 92 were target items, and 104 were filler items. The target items were 46 Dutch-English cognate items and 46 English non-cognate items. An item was considered a cognate if it had the same meaning in English and Dutch and the Levenshtein distance [15] (not considering word stress) between the Dutch and the English pronunciations was 5 or less (mean 3.3). The cognates and non-cognates had similar log subtitle word frequencies (SUBTLWF, [16]; mean frequency for cognates and non-cognates: 2.18 and 2.41, respectively; t-test: t = -1.68, p = 0.1). They were all trisyllabic and had a schwa in the second syllable. Main stress was on the first syllable, whereas it was on the final syllable in the cognates' Dutch equivalents (e.g., English /'ɪmpətənt/ versus Dutch /ɪmpo'tɛnt/ impotent).

The stimuli were recorded by a male native speaker of British English. All target words were recorded in full form and without the schwa in the second syllable. There were two groups of participants: 31 students (mean age = 22.1 years, SD = 2.3) of Radboud University, all native speakers of Dutch and master students of English-taught degrees. They were highly proficient in English as evidenced by their scores on the LexTALE proficiency task (mean = .76, SD = .11; [17]). The native group comprised 38 students (mean age = 21.5 years, SD = 3.2) of the University of Cambridge who did not speak Dutch.

The `lmer` model for the 'English' data is shown in Table 1. The significance patterns are clearly dependent on the RT measure used as dependent variable. The most obvious difference is the change in sign of the β for logwdur (log(word duration)), which is positive in the case RT from onset and negative in the case RT from offset. Table 1 also shows that the effects of Frequency and Cognate status are different when measured from

Table 1: *LMER model for the English data*

English data			
	Estimate	Std. Error	t value
(Intercept)	4.434e+00	9.745e-01	4.550
offset	2.250e+00	9.222e-01	2.440
cogn	-2.488e-02	3.508e-02	-0.709
reduced	-2.999e-02	2.310e-02	-1.298
lnFreq	-6.278e-02	2.403e-02	-2.613
lnwrddur	3.473e-01	1.250e-01	2.778
Native	-4.815e-01	7.225e-01	-0.666
Trialnr	5.336e-05	6.559e-05	0.814
maRT	6.457e-02	7.364e-02	0.877
prevBVis	4.594e-03	2.513e-03	1.828
cogn:Native	2.497e-02	3.052e-02	0.818
reduced:Native	3.076e-02	3.001e-02	1.025
lnFreq:Native	4.372e-02	2.093e-02	2.089
lnwrddur:Native	3.644e-02	1.097e-01	0.332
offset:cogn	-5.198e-02	3.304e-02	-1.573
offset:reduced	-1.337e-01	3.257e-02	-4.105
offset:lnFreq	-1.253e-01	2.266e-02	-5.529
offset:lnwrddur	-1.664e+00	1.181e-01	-14.090
offset:Native	1.695e+00	1.017e+00	1.666
offset:Trialnr	2.924e-04	9.173e-05	3.188
offset:maRT	1.110e+00	7.080e-02	15.675
offset:prevBVis	3.171e-03	3.506e-03	0.904
offset:cogn:Native	8.231e-02	4.301e-02	1.914
offset:reduced:Native	4.552e-02	4.236e-02	1.074
offset:lnFreq:Native	5.924e-02	2.948e-02	2.010
offset:lnwrddur:Native	-3.228e-01	1.546e-01	-2.088

Table 2: *LMER model for the French data*

French data			
	Estimate	Std. Error	t value
(Intercept)	8.653e+00	2.695e-01	32.103
offset	-1.835e+00	1.885e-01	-9.735
Native	-8.775e-01	1.362e-01	-6.442
lnwdur	1.011e-01	3.094e-02	3.267
lnFreq	-2.580e-02	3.946e-03	-6.539
reduced	-8.118e-02	1.020e-02	-7.960
cogn	3.259e-02	1.509e-02	2.160
maRT	-2.563e-01	2.322e-02	-11.038
prevBVis	2.201e-03	6.522e-04	3.375
blockNr	-7.416e-04	3.256e-03	-0.228
Native:lnwdur	9.096e-02	1.976e-02	4.603
Native:lnFreq	2.480e-02	2.362e-03	10.500
Native:reduced	-1.994e-02	1.312e-02	-1.520
Native:cogn	-5.376e-03	9.107e-03	-0.590
offset:Native	1.099e+00	1.851e-01	5.934
offset:lnwdur	-9.205e-01	1.928e-02	-47.732
offset:lnFreq	-2.304e-02	2.325e-03	-9.911
offset:reduced	2.394e-02	1.273e-02	1.880
offset:cogn	3.904e-02	8.909e-03	4.382
offset:maRT	9.681e-01	1.860e-02	52.057
offset:prevBVis	3.657e-03	9.073e-04	4.030
offset:blockNr	2.193e-02	3.101e-03	7.073
offset:Native:lnwdur	-1.801e-01	2.789e-02	-6.457
offset:Native:lnFreq	2.525e-02	3.330e-03	7.582
offset:Native:reduced	-5.131e-02	1.853e-02	-2.769
offset:Native:cogn	-1.035e-03	1.284e-02	-0.081

offset and onset. This suggests that the processing of the English stimuli differed between native and non-native listeners. Further examination of the data then needs to show whether the same processes are just delayed (e.g., a mere stronger or even offset-only effect of word frequency may indicate that lexical selection and activation might be delayed), or that there are different processes at play (e.g., a different sign of the β of word frequency). Also, the significant interaction of cognate status with group and `offset` (0/1) could reveal when and how semantic co-activation of L1 affects the processing of non-natives.

3.2. An experiment with Dutch learners of French

The second set of data comes from [9]. Their stimuli consisted of 44 French, morphologically simple, bisyllabic target words with a schwa in the first syllable (e.g., 'le menu') and 520 fillers. Of the fillers, 44 were pseudowords with schwa in the first syllable (e.g., 'le beseuil'). The bisyllabic targets and pseudowords were recorded in their full form, with schwa, and without the schwa. There were two groups of participants. One group consisted of 47 Dutch undergraduate students of French (aged 19–30 years; 11 males). All were born and raised in the Netherlands, had taken French classes for five or six years at secondary school, and had studied French at university for at least seven months and at most three years and seven months. Their CEFR-levels corresponded to C1–C2 level [18]. The second group comprised 36 native speakers of French from Paris (three males), aged between 19 and 30 years. The `lmer` model of the log-transformed RT data of [9] are presented in Table 2. The predictor `log(freq)` is the logarithm of the counts of the number of occurrences in the film subtitles in the *Lexique* data base [19]. The frequencies of identically spelled word forms with different word class were simply added. Although the concept 'cognate' was not included explicitly in the experimental design, we marked a subset of the word stimuli as cognates for this study to enable a comparison with the data from [8]. Before computing $\log(RT)$ the RT sequences were filtered using the methods presented in [14]. In computing $\log(\text{RT}_{\text{offset}})$ the stimuli for which the response was given before the offset of the stimulus were discarded. Both models contained the following simple random structure: (1|word) + (1|subject).

Table 2 shows a clear difference with respect to significance patterns depending on from which time point RTs are measured, the two most obvious differences being the change in sign of the β for logwdur (log(word duration)), which is positive in the case RT from onset and negative in the case RT from offset, and the change in sign of the β for reduction, which is negative in the case RT from onset and positive in the case RT from offset. Table 2 further demonstrates that cognate status had a different effect when measured from onset or offset, in both the native and non-native data. This does not mean that the French natives were sensitive to the Dutch representations, as they had no knowlegde of Dutch, but it may show that cognate are "special items" in the sense that these items might not have many neighbors [20] that hamper or facilitate the processing. Finally, similar to the English data, log Frequency plays a significantly different role during word recognition for native versus non-native speakers of French.

3.3. BALDEY

Table 3 presents the `lmer` model of the log-transformed RT data of BALDEY ([1]). BALDEY contains about 110,000 reaction times from a large-scale Dutch lexical decision experiment. The table shows the model results pertaining to the fixed structure of the `lmer` models. The presented model was obtained

Table 3: *LMER model for the Dutch data.*

Dutch data			
	Estimate	Std. Error	t value
(Intercept)	1.868e+00	5.714e-01	3.270
offset	-2.597e+00	7.692e-02	-33.768
lnwdur	2.943e-01	1.028e-02	28.631
lnFreq	5.739e-02	1.917e-02	2.994
classnom	5.997e-03	5.392e-03	1.112
classverb	3.221e-02	5.369e-03	6.000
maRT	4.760e-01	7.529e-02	6.322
prevBVis	2.853e-03	4.624e-04	6.169
compnd_A+N	-2.835e-02	1.609e-02	-1.762
compnd_N+A	-2.951e-03	1.845e-02	-0.160
compnd_N+N	2.603e-02	6.961e-03	3.740
session	2.223e-03	3.355e-04	6.627
trial	2.338e-05	5.602e-06	4.173
lnwdur:lnFreq	-1.049e-02	2.993e-03	-3.504
offset:lnwdur	-8.309e-01	9.534e-03	-87.150
offset:lnFreq	3.601e-02	1.778e-02	2.026
offset:classnom	3.605e-03	4.981e-03	0.724
offset:classverb	5.052e-02	4.947e-03	10.213
offset:maRT	9.944e-01	6.723e-03	147.913
offset:prevBVis	3.114e-03	6.465e-04	4.817
offset:compnd_A+N	-4.412e-02	1.493e-02	-2.956
offset:compnd_N+A	-2.592e-02	1.726e-02	-1.502
offset:compnd_N+N	-1.153e-02	6.520e-03	-1.768
offset:lnwdur:lnFreq	-7.257e-03	2.778e-03	-2.612

using the recommendations in [21] with respect to inclusion of fixed terms, interactions and random slopes. As a result, `maRT` was used as slope under participant in the random structure: (1|word) + (1|subject) + (0+maRT|subject)

Addition of more random slopes yielded models that did not converge. Both models (onset and offset) were estimated on 95613 data points, and checked via a criticism phase in which outlier RT data with a residual of more than 3 standard deviations away from the grand mean were removed. Table 3 shows the resulting `lmer()` model on BALDEY for RT from onset (top) and RT from offset (bottom).

The table shows a clear difference with respect to significance patterns depending on from which time point RTs are measured, the most obvious difference being the change in sign of the β for logwdur (log(word duration)), which is positive in the case RT from onset and negative in the case RT from offset.

It can be seen that nearly all interaction of `offset` with predictors of interest, in particular word duration and log Frequency, are significant. This shows that RT_{onset} and RT_{offset} models provide significantly different windows on the RT structure.

4. Discussion and conclusion

We investigated whether two different operational definitions of reaction time (RT), RT_{onset} and RT_{offset}, reveal different roles of predictors in `lmer` models. We analyzed the data of three different lexical decision experiments with native listeners or a combination of native and non-native listeners. For these data, we built a model that combines RT_{onset} and RT_{offset} as the dependent variable and an additional factor with levels `onset` and `offset`. In all three data sets we find significant interactions between that factor and several predictors that are of theoretical interest. This suggests that these predictors behave significantly different in models of onset or offset RTs, and therefore that analyses using RT_{onset} or RT_{offset} do indeed focus on substantially different aspects of reaction time in these type of experiments.

In all three data sets the interaction between onset/offset and stimulus duration was among the ones with the largest t-value. Separate analyses showed that the β for this predictor is positive with RT_{onset}, and negative for RT_{offset}. Thus, the predictor that tends to explain most of the variance in RTs [1, 2, 3] appears to play a different role, depending on whether RT is measured from stimulus onset or offset. Most probably, with RT_{offset} longer stimuli profit from a larger amount of information accumulated while the stimulus unfolds. In the English and French data there is a significant three-way interaction that shows that native listeners benefit more from the information collected during the course of the stimulus.

The predictor `frequency` (always log-transformed) yields seemingly complex results. In the BALDEY data, where a large proportion of the stimuli are morphologically complex, the β of `lnFreq` and `offset:lnFreq` are positive, but this counter-intuitive effect is compensated by negative βs of the interactions `lnwdur:lnFreq` and `offset:lnwdur:lnFreq`. In the English and French data, where all stimuli are monomorphematic, we see a strong negative β for `lnFreq`, but combined with similarly surprising positive βs of the interactions `Native:lnFreq` and `offset:Native:lnFreq`. This implies that the non-native listeners benefit more from lexical frequency than natives, especially after stimulus offset. While it is tempting to explain this non-native advantage in terms of lexical access, it should be realized that this explanation is negated by the inverse effect of lexical frequency in natives.

The predictor reduction played a significantly different role in the processing of natives and non-natives in the French data, but not so much in the English data. This may be due to the differences in prosodic structure between English and French. Whereas English words are reduced in post-stress position, French words are reduced in prestress position. Apparently, the relative difficulty in processing a word depends on the complexity of the onset of the word, where reduction has a large impact. If the first syllable is intact (i.e., without reduction), non-native listeners may not experience more problems in subsequent processing than natives do. Different effects of prestress and post-stress reduction are in line with [22, 23].

Finally, cognate status appeared to be more important in the French data, where this factor was not part of the original design. More research is needed to understand its role, especially in auditory lexical decision experiments, where the definition of 'cognate ´ is more difficult than in visual experiments.

We conclude that a comparison between RT_{offset} and RT_{onset} models provide valuable insights into the way in which the predictors of interest influence the processing of stimuli. More specifically, an interaction of a given predictor and the predictor 'offset' may reveal a change over time of the potential time courses of cognitive processes that are assumed to be reflected by these predictors. A comparison of offset and onset models is especially useful when comparing different groups of listeners, for instance native and non-native listeners, for whom differences in processing speed are to be expected.

5. Acknowledgements

The English and French data were collected within an ERC Consolidator grant from the European Research Council [grant number 284108] awarded to Mirjam Ernestus.

6. References

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

[2] L. Ferrand, A. Méot, E. Spinelli, B. New, and C. Pallier, "MEGALEX: A megastudy of visual and auditory word recognition," *Behavior Research Methods*, vol. 50, no. 3, pp. 1285–1307, 2018.

[3] B. V. Tucker, D. Brenner, D. K. Danielson, M. C. Kelley, F. Nenadić, and M. Sims, "The massive auditory lexical decision (mald) database," *Behavior Research Methods*, vol. 51, no. 3, pp. 1187–1204, 2019.

[4] V. Taler, G. Aaron, L. Steinmetz, and D. Pisoni, "Lexical neighborhood density effects on spoken word recognition and production in healthy aging." *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, vol. 65, no. 5, pp. 551–560, 2010.

[5] B. Munson, C. Swenson, and S. Manthei, "Lexical and phonological organization in children," *Journal of Speech, Language, and Hearing Research*, vol. 48, no. 1, pp. 108–124, 2005.

[6] 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.

[7] S. Dufour, A. Brunellière, and U. H. Frauenfelder, "Tracking the time course of word-frequency effects in auditory word recognition with event-related potentials," *Cognitive Science*, vol. 37, no. 3, pp. 489–507, 2013.

[8] 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.

[9] S. Brand and M. Ernestus, "Listeners' processing of a given reduced word pronunciation variant directly reflects their exposure to this variant: Evidence from native listeners and learners of french," *Quarterly Journal of Experimental Psychology*, vol. 71, no. 5, pp. 1240–1259, 2018.

[10] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Selected Papers of Hirotugu Akaike*, P. E., T. K., and K. G., Eds. New York: Springer, 1998, springer Series in Statistics (Perspectives in Statistics).

[11] G. E. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, no. 2, pp. 707–710, 1978.

[12] D. Lakens and A. R. Caldwell, "Simulation-based power-analysis for factorial anova designs," May 2019. [Online]. Available: psyarxiv.com/baxsf

[13] R. L. Wasserstein, A. L. Schirm, and N. A. Lazar, "Moving to a world beyond "$p < 0.05$"," *The American Statistician*, vol. 73, no. sup1, pp. 1–19, 2019.

[14] 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.

[15] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals." *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, 1966.

[16] M. Brysbaert and B. New, "Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English," *Behaviour Research Methods*, vol. 41, pp. 977–990, 2009.

[17] K. Lemhöfer and M. Broersma, "Introducing LexTALE: A quick and valid lexical test for advanced learners of english," *Behavior Research Methods*, vol. 44, no. 2, pp. 325–343, 2012.

[18] Council of Europe, "Common european framework of reference for languages: Learning, teaching, assessment," 2011. [Online]. Available: http://www.coe.int/t/dg4/linguistic/cadre1_en.asp

[19] B. New, C. Pallier, M. Brysbaert, and L. Ferrand, "Lexique 2 : A new French lexical database," *Behavior Research Methods, Instruments, & Computers*, vol. 36, no. 3, pp. 516 – 524, 2004.

[20] K. Mulder, T. Dijkstra, R. Schreuder, and H. R. Baayen, "Effects of primary and secondary morphological family size in monolingual and bilingual word recognition," *Journal of Memory and Language*, vol. 72, pp. 59–84, 2014.

[21] H. Matuschek, R. Kliegl, S. Vasishth, and D. Baayen, H.and Bates, "Balancing type I error and power in linear mixed models," *Journal of Memory and Language*, vol. 94, pp. 305 – 315, 2017.

[22] A. Bürki and M. Gaskell, "Lexical representation of schwa words: Two mackerels, but only one salami," *Journal of Experimental Psychology: Learning Memory and Cognition*, vol. 38, no. 3, pp. 617 – 631, 2012.

[23] P. LoCasto and C. Connine, "Rule-governed missing information in spoken word recognition: Schwa vowel deletion," *Perception & Psychophysics*, vol. 64, pp. 208 – 219, 2002.