INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Analyzing reaction time and error sequences in lexical decision experiments

L. ten Bosch[1], *L. Boves*[1], *K. Mulder*[1]

[1]Radboud University Nijmegen, The Netherlands
{l.tenbosch, l.boves, k.mulder}@let.ru.nl

Abstract

Reaction times (RTs) are used widely in psychological and psycholinguistic research as inexpensive measures of underlying cognitive processes. However, inferring cognitive processes from RTs is hampered by the fact that actual responses are the result of multiple factors, many of which may not be related to the process of interest. In lexical decision experiments, the use of RTs is further complicated by the fact that the response to some stimuli is missing, and the fact that part of the responses are 'incorrect'.
In this paper we investigate the distribution of missing and incorrect responses in the RT sequences of two large lexical decision experiments. It appears that a substantial part of incorrect responses cluster together. Then, we investigate the effect of clusters of incorrect responses on surrounding RTs.
Also, we extend previous research on methods for discovering and removing so-called local speed effects from RT sequences. For this purpose, we show that a recently introduced graph-based RT analysis method can help to better understand and analyze RT sequences.
Index Terms: reaction time, visibility graph, local speed effect, error distribution, lexical decision.

1. Introduction

Despite the emergence of eye tracking and brain imaging techniques such as EEG, MEG, fMRI and fNIRS, behavorial experiment that rely on reaction time (RT) measurements remain an important technique for investigating cognitive processes in psychological and psycholinguistic research. While RTs are easy and inexpensive to measure, the statistical processing of RTs with the goal to elucidate underlying cognitive processes is fraught with problems, the importance of which is easy to underestimate. Observed RTs (e.g., measured via a button press) are the result of a number of sequential and parallel cognitive, neuro-physiological and mechanical processes [1, 2]. Confounding factors include long-term effects (participant's health condition, age, gender, handedness, general cognitive abilities, gaming experience, etc. [3]) and medium-term effects (attention fluctuation, strategy changes, fatigue). Medium-term effects are collectively referred to as 'local speed effects'.

Distributions of RTs are notoriously non-normal. It is possible to apply transformations, such as a log-transform, to make the distributions more normal, but these procedures require arbitrary decisions, e.g., for discovering outliers. The use of non-Gaussian distributions, such as the ex-Gaussian distribution, avoids the need for removing 'exceptionally large' RT values. Research using ex-Gaussian distributions has shown that the skewness parameter τ tends to carry more information about differences between persons, groups or conditions than the conventional mean and variance of the distributions [4].

Reaction times in psycholinguistic experiments, such as lexical decision experiments, come in sequences. Treating the RT values as an unordered set loses essential information. Perhaps the most compelling indication of the importance of treating RTs as ordered sequences is the predictive power of a predictor 'previous RT' in statistical regression models (e.g., [5]).

Another vexing problem in treating RTs in lexical decision experiments is how to handle 'incorrect' decisions. An incorrect answer might provide useful information: if a participant simply did not know a real word, a possibility that seems to be quite real in experiments addressing the processing of words in a second language, a 'nonword' response might very well reflect a valid representation of the cognitive processes under investigation.

At the same time 'incorrect' responses may be truly errors, perhaps caused by a lack of attention. Usually these errors are removed from the data for analysis. However, attention fluctuations are usually considered as one of the 'local speed effects' that are likely to affect a (short) sequence of stimuli. This suggests that it is potentially relevant to investigate the distribution of errors in a session.

In this paper we study RT sequences by looking at the distribution of between-error lags in combination with a `natural visibility graph` (NVG) [6], a recently developed analysis tool based on graph theory. The structure of RT sequences can be characterized and further analyzed by converting them to NVGs. Software for using NVGs is provided in [7], which builds on a widely used Python package for constructing and analyzing graphs `networkX` [8]. One factor that almost certainly affects RTs is attention. NVGs have been used to infer attention fluctuations in the diagnosis of ADHD and dementia-related conditions [9].

In this paper we use RTs and accuracies of two large-scale lexical decision experiments [10, 11] to investigate the distributions of errors and the options offered by the newly introduced natural visibility graphs.

2. The databases

The BALDEY corpus [10] contains RTs related to lexical decisions by twenty native Dutch listeners (10 male, 10 female, 18 to 23 years) without reported hearing problems; participants were paid to participate in this experiment. 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.

The second database was collected by [11] in an experiment that investigated several properties of words in a second language lexical decision experiment. Forty right-handed non-native listeners of English (mean age = 20.9 years, SD = 2.2) participated in the experiment. All were native speakers of Dutch and master students of English-taught degrees at Radboud University. They were highly proficient in English as evidenced by their scores on the LexTALE proficiency task (mean = .83, SD = .37; [12]. For several reasons five participants were discarded. The experiment involved in total 900 stimuli, half

http://dx.doi.org/10.21437/Interspeech.2019-2611

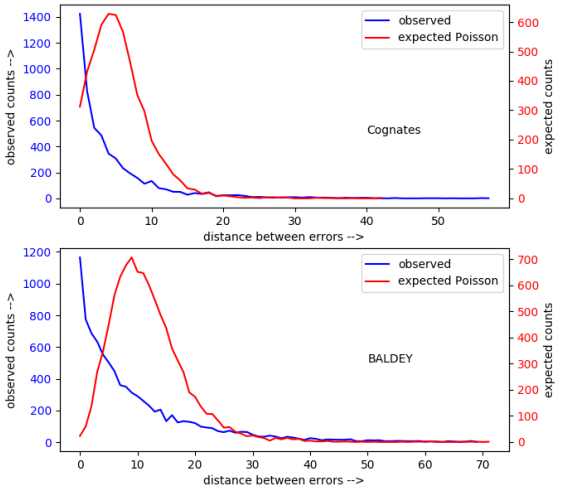

Figure 1: *Comparison between observed distributions of between-error distances and predictions based on Poisson distributions. Top panel: Cognate database; bottom panel: BALDEY.*

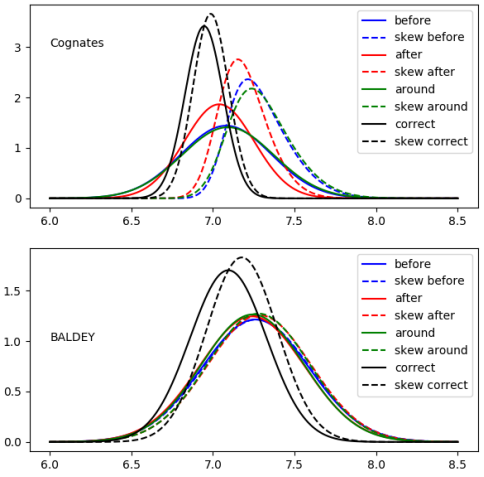

Figure 2: *Distributions of logRT values around errors and in error-free stretches.*

of which were pseudo words. The target items were 68 Dutch-English cognate items that contained a schwa in pre-stress position and 46 cognates that contained a schwa in post-stress position, in addition to 68 and 46 non-cognate items, respectively. An item was considered a cognate if it had the same meaning in English and Dutch and the Levenshtein distance (not considering word stress) between the Dutch and the English pronunciations was 5 or less (mean 3.71 for the pre-stress stimuli, and 3.3 for the post-stress stimuli [13, 11]). The cognates and non-cognates had similar log subtitle word frequencies (SUBTLWF [14]; mean log frequency for cognates and non-cognates in the pre-stress condition: 2.25 and 2.08, respectively, and in the post-stress condition 2.18 and 2.41, respectively. The 222 filler items were disyllabic, tri-syllabic or four-syllabic real words with the position of word stress varying between words. These items were matched to the experimental set on number of syllables and frequency of occurrence. The 450 pseudo words were generated by means of Wuggy [15] on the basis of the target and filler words. We will refer to this data as the Cognates database. The pre-stress and post-stress targets and their matched controls were presented in two separate blocks. The order of the blocks was counterbalanced.

3. Error distributions

If most or all discrepancies between the word/nonword status assigned to a stimulus by the experimenter and the participant's decision are incidental, the distances between successive events (which we call 'errors' for convenience of formulation) would follow a Poisson distribution. We calculated the observed distribution of the between-error distances for the BALDEY and the Cognates databases. We also simulated Poisson-distributed distances for each experimental session based on the observed number of errors in that session and the number of stimuli in the session. The results are summarized in Figure 1. The top panel of Fig. 1 shows the result for the Cognate data base, the bottom panel for the BALDEY data base. The mismatch between the observed (blue) and the simulated (red) distributions is obvious, and similar for both data bases. The large number of counts of distance=0 in the observed distributions is due to the occurrence of clusters of erroneous decisions. These clusters are virtually absent from the simulations that assume a Poisson distribution. Thus, we must conclude that the distribution of the errors in a session is not random.

The total number of erroneous decisions in BALDEY is 9335 (8.4%), 7277 of which can be considered as isolated or incidental errors, in the sense that the decision on the preceding and following stimuli were correct. In BALDEY we found 749 sequences of two consecutive errors, 108 sequences of three, 25 sequences of four five sequences of five, and one sequence of six and of seven errors. In session number 10 of participant number 18 there are clusters of consecutive errors of lengths 12, 18, 23 and 45. There is a weak trend for the number of error clusters to grow with the session number.

In the Cognates data base there are 5534 errors, 3366 of which are isolated. We found 268 error sequences of length 2; 56 of length 3; 18 of length 4; 7 of length 5, 5 of length 6, 4 of length 7; 2 of length 8; and 1 of length 10. In addition, pp32 had 3 stretches of 11 contiguous errors, two of 12, one of 14; one of 16; one of 23; one of 40 and one of 46.

To see whether logRTs around clusters of errors differ from logRTs in intervals without decision errors we computed the average RT of a window of five stimuli just before the clusters, around the middle of the clusters and after the clusters, and compared the distributions to averages of five consecutive logRTs where no errors occurred. The results are summarizes in Fig. 2. In both databases we see similar trends: RTs tend to be longer around clustered errors than in stretches without errors. Also, the standard deviations of the RTs around errors are larger than in error-free stretches. Especially in the Cognates database there is s stronger trend for the distributions to be more skewed to the right than in the BALDEY data.

4. Natural Visibility Graphs

In [5] we investigated signal processing and time series analysis techniques to come to grips with local speed effects in RT sequences. Here, we explore another technique for analyzing

sequences of RT values, based on graph theory. In [16] it is explained how time series can be converted to graphs, and how graph theory can be used to expose structure in a time series. Graphs are mathematical constructs designed to model pairwise relations between objects. From sequences of RT values we can construct a specific type of graph, namely Natural Visibility Graphs (NVG).

This construction of a NVG is explained in Figure 3. A sequence of RTs is represented as an array of vertical poles, arranged along a straight line, at equal distances. Each RT is associated to one pole, and the RT itself specifies the height of its pole. This pole array is the basis of the graph construction. Each RT is *connected* to all RTs that can be 'seen', according to the arrangement of the poles. Each resulting connection is an arc in the visibility graph. In the figure the first RT can 'see' the second and the third, but not the fourth, because the third blocks the view. Evidently each RT can always see its left and right neighbor, because there is nothing that can block the visibility of the direct neighbors. The Python script published by [7] converts a time series into the specification of a graph. The processing of the graph is done using `networkX` [6].

NVGs formed from an RT sequence can be used to characterize the complete time series. In the absence of clear local speed effects, and under the assumption that the sequence of stimuli is randomly distributed, one would expect that all nodes have about the same degree: the differences between successive RT values are similar in all segments of the graph. However, local speed effects would introduce local larger variations in the degree of the nodes. This links basic aspects of RT sequences as graphs with the idea elaborated in [5] that the spectrum of an 'ideal' RT sequence must be flat.

One characteristic of a graph that can be obtained with `networkX` is the so called `degree` of the nodes, i.e., the number of nodes to which a specific node is connected. In [9] node degree distributions were used to identify sessions (actually: participants) with 'exceptional' node degree distributions, to see if outlier distributions predict attention disorders (i.c., ADHD). We applied a similar analysis to the node degree distributions of all (sub-)sessions in the Cognates and BALDEY databases. While there is some variance in the set of distributions, for example characterized by tuples such as (mean, skewness), we did not find reliable graph-related indicators for (sub-)sessions that might be problematic in terms of RT skewness. Even sessions with large numbers of incorrect decisions did not stand out. In the BALDEY data the node distributions of all data collapsed over each of the ten sessions did not replicate the finding that the number of error clusters becomes slightly larger in the later sessions.

However, the node degrees do contain information that can be exploited in regression models. We modified the definition of the degree of a node. In the original definition, the degree is defined by counting the arcs to visible nodes in both forward and backward direction. Here, we use the 'backward looking degree' in the previous node as a predictor for the current RT (i.e., the number of *preceding* RT(t-n-1) values that are visible from RT(t-1), exactly similar to the use of the 'previous RT' in an `lmer` model in `R` that predicts log(RT) in the Cognate data:

$$\begin{aligned}\text{lmer}(logRT \sim prevlogRT + log(duration) + task \\ +Correct + logFreq + nextCorrect + prevBVis \\ +(1|word) + (1|PPN)),\end{aligned} \quad (1)$$

with `prevBVis` representing the backward-looking visibility degree of the previous RT. In table 1 it can be seen that `prevBVis` made a small, but very significant contribution to the prediction accuracy. Apparently, this predictor contains information about effects of previous stimuli that are not easy to capture otherwise. In [5] it was shown that there is a longer stretch (longer than 1) of preceding stimuli that can affect the present one. However, so far we have only been able to use that knowledge by applying a linear filter, the operation of which is independent of the idiosyncrasies of specific stimulus sequences. The backward-looking node degree will make that 'filter' stimulus-dependent in a well-defined graph-based nonlinear fashion.

Table 1: *Result of the mixed effects regression model 1; prevBVis is one of the highly significant predictors.*

```
Random effects:
 Groups   Name        Variance Std.Dev.
 word     (Intercept) 0.003451 0.05874
 PPN      (Intercept) 0.008597 0.09272
 Residual             0.044429 0.21078
Number of obs: 30963, groups: word, 907;
PPN, 35

Fixed effects:
                Estimate Std.Err  t value
(Intercept)     4.701338  0.078033  60.248
myprevlogRT     0.184327  0.005976  30.844
log(duration)   0.168802  0.010062  16.776
taskprestress   0.019879  0.004696   4.233
Correct        -0.032691  0.001850 -17.666
mylogFreq      -0.041786  0.001455 -28.718
nextCorrect     0.003382  0.001717   1.970
prevBVis       -0.005619  0.000533 -10.537
```

The fact that the node degrees contain additional information about the RT sequences is confirmed by the fact that on average the node degree sequences explain at best some 20% of the variance in the log(RT) sequences. We found a similar contribution of `prevBVis` in a regression model that predicts logRTs in the BALDEY database.

Node degree distributions can also be used to discover unexpected behavior of specific stimulus types. An example is shown in Figure 4, which shows the distributions of the node degrees for the stimulus types in the Cognates experiment. Input for the construction of the graph were the raw RT values minus the duration of the stimulus words. No transformations (such as a log-transform to make distributions more normal) were applied, and no data were discarded. The most striking information in the figure is the large difference between the distributions related to the reduced words (both controls and cognates) in the *prestress* condition. The degree of these stimuli is on average larger than of all other conditions, and the standard deviations are also among the largest.

We performed the same analysis on the BALDEY data, looking for differences in the node degree distributions of adjectives, nouns and verbs, as well as nonwords with the morphological structure of adjectives, nouns and verbs. There appeared to be a small difference between verbs (both words and nonwords) on the one hand and adjectives and nouns on the other, but that difference was nowhere near as striking as the outstanding position of the prestress reduced words in the Cognate data.

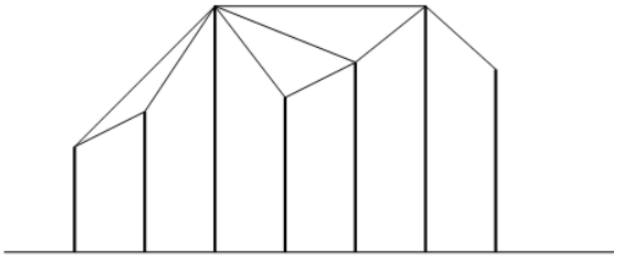

Figure 3: *Example of the conversion of an RT sequence to a natural visibility graph.*

5. Discussion

5.1. Error distributions

In both databases we see that a substantial proportion of the decision errors come in clusters, mostly of two or three consecutive errors. We also see that the logRT values around clustered errors tend to be larger than in error-free stretches. It is tempting to speculate that clustered errors in combination with longer logRTs suggest that the participants' attention was decreasing around the time of the presentation of the stimuli in error clusters. That holds especially for the clusters that are preceded by relatively slow responses. However, with only the behavioral RT data, it is more dangerous to make conclusions about the cause of slower responses *after* a cluster of errors. There is at least one other explanation for the occurrence of longer RTs after a cluster: The cause might be the first error in a sequence; participants may have been unsure about the word/nonword status of that stimulus, and being conscious of that uncertainty might have affected the processing of one or two subsequent stimuli. It can be argued that a similar effect might also occur when the decision on the 'difficult' happens to be correct. However, With only RT data it is not evident how we can discover stimuli with a correct response that still have a large effect on subsequent stimuli because a participants' attention is caught by doubts about the correctness of a recent decision.

5.2. Natural Visibility Graphs

Natural Visibility Graphs were introduced as a potentially powerful means for discovering structure in time series, and for discovering differences between sets of time series. In [9] NVGs were used successfully to distinguish RT sequences of eight year old children diagnosed with ADHD in a visual lexical decision task from the sequences of controls. We started our investigation of NVG-based analysis of RT sequences from lexical decision experiments with the aim to develop a new method for discovering and removing local speed effects. The idea was that local speed effects are most likely different between experimental sessions, and that these difference would lead to differences between node degree distributions. While there are indeed differences between the node degree distributions of different sessions, these differences are far smaller and far more similar to Gaussian distributions than in the data in [9]. Apparently, university students who volunteer to participate in lexical decision experiments form a much more homogeneous population than a group of eight-year-olds that includes a few ADHD children. Even the node degree distributions of RT sequences with a relatively high proportion of erroneous decisions did not stand out.

However, it appeared that node degree distributions can contain useful information, of which many details are not yet fully understood. In the Cognates data node degree distributions showed a clear difference between stimuli with a heavily reduced pre-stress syllable and all other stimulus types (including pseudo words). Also, while distributions of node degrees covering a complete session may not be very illuminating, it appeared that local variations in node degree make a significant contribution in predicting upcoming RTs. This does corroborate our assumption that NVGs can play a role in reducing the impact of local speed variations in the analysis of RT data in lexical decision experiments.

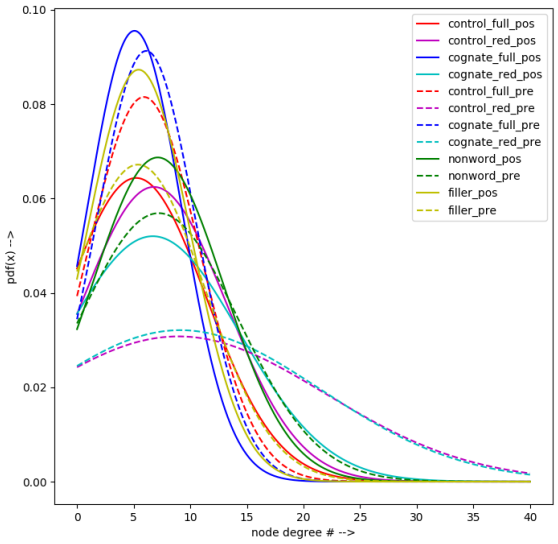

Figure 4: *Distributions of the degree distributions of RT sequences related to different types of stimuli in a lexical decision experiment after conversion of the sequences to visibility graphs.*

6. Conclusions

In this paper we show that a substantial part of erroneous responses in lexical decision experiments come in clusters. The distributions of the distance between errors in two large databases from lexical decision experiments were clearly different from the distributions one would expect if errors would follow a Poisson distribution. In addition, it appeared that RT values around clusters of errors tend to be longer and more variable. This warrants a more in-depth analysis of the impact of the way erroneous decisions are handled in statistical analyses of lexical decision experiments.

We also showed that recently proposed Natural Visibility Graphs can be used to remove confounding local speed effects from RTs in lexical decision experiments.

7. Acknowledgements

This work was supported by an ERC consolidator grant (284108) to Mirjam Ernestus. Also, she was involved in the design of the Cognates project.

8. References

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

[2] S. Sternberg and B. Backus, "Sequential processes and the shapes of reaction time distributions," *Psychological Review*, vol. 122, no. 4, pp. 830–837, 2015.

[3] J. J. Lee and C. F. Chabris, "General cognitive ability and the psychological refractory period: Individual differences in the minds bottleneck," *Psychological Science*, vol. 24, no. 7, pp. 1226 – 1233, 2013.

[4] S.-L. H. Gu, S. S.-F. Gau, S.-W. Tzang, and W.-Y. Hsu, "The ex-Gaussian distribution of reaction times in adolescents with attention-deficit/hyperactivity disorder," *Research in Developmental Disabilities*, vol. 34, no. 11, pp. 3709 – 3719, 2013.

[5] 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. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1728

[6] L. Lacasa, B. Luque, F. Ballesteros, J. Luque, and J. C. Nuño, "From time series to complex networks: The visibility graph," *Proceedings of the National Academy of Sciences*, vol. 105, no. 13, pp. 4972–4975, 2008.

[7] G. Iacobello, S. Scarsoglio, and L. Ridolfi, "Visibility graph analysis of wall turbulence time-series," *Physics Letters A*, vol. 382, no. 1, pp. 1 – 11, 2018.

[8] A. A. Hagberg, D. A. Schult, and P. J. Swart, "Exploring network structure, dynamics, and function using NetworkX," in *Proceedings of the 7th Python in Science Conference (SciPy2008)*, G. Varoquaux, T. Vaught, and J. Millman, Eds., 2008, pp. 11 – 15.

[9] A. Mira-Iglesias, J. A. Conejero, and E. Navarro-Pardo, "Natural visibility graphs for diagnosing attention deficit hyperactivity disorder (ADHD)," *Electronic Notes in Discrete Mathematics*, vol. 54, pp. 337 – 342, 2016, discrete Mathematics Days - JMDA16.

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

[11] K. Mulder, L. Wloch, and M. Ernestus, "Effects of cognate status and location of reduced syllables on lexical decisions of Dutch-English bilinguals," submitted.

[12] 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.

[13] 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.

[14] 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.

[15] E. Keuleers and M. Brysbaert, "Wuggy: A multilingual pseudoword generator," *Behavior Research Methods*, vol. 42, pp. 627–633, 2010. [Online]. Available: https://doi.org/10.3758/BRM.42.3.627

[16] S. Yan and D. Wang, "Time series analysis based on visibility graph theory," in *7th International Conference on Intelligent Human-Machine Systems and Cybernetics*, vol. 2, 2015, pp. 311–314.