Non-native Word Recognition in Noise: The Role of Word-initial and Word-final Information

Juul Coumans[1,2], Roeland van Hout[1], and Odette Scharenborg[1,3]

[1] Centre for Language Studies, Radboud University Nijmegen, The Netherlands
[2] IMPRS for Language Sciences, Nijmegen, The Netherlands
[3] Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, The Netherlands
{J.Coumans, R.vanHout, O.Scharenborg}@let.ru.nl

Abstract

When listening in noisy conditions, word recognition seems to be much harder in a non-native language than in one's native language. Native listeners use both word-initial and word-final information for word recognition in clean listening conditions, where word-initial information is the most important, though word-final information becomes relatively more important when listening in noise. This study investigates whether non-native listeners are able to use word-initial and word-final information when recognizing words in noise, and whether these information sources are equally important when listening conditions become increasingly harder. Forty-seven Dutch students participated in an English word recognition experiment, where either a word's onset or offset was masked by speech-shaped noise with different signal-to-noise ratios. The results showed that non-native listeners are able to use both word-initial and word-final information for word recognition, but fewer words were recognized with increasing difficulty of the listening conditions when the onset of words was masked. Thus, word-initial information is more important than word-final information for word recognition when listening conditions become harder. This increasing effect occurred independently from the proficiency level in the non-native language of the participants, although proficiency level was correlated to test performance in general.

Index Terms: non-native spoken word recognition, listening in noise, proficiency.

1. Introduction

In everyday life, the conditions for recognizing speech can be difficult. Traffic, loud music, and so on are all sources of noise, which may hinder speech recognition. Listening in the presence of noise is hard in one's native language but even more challenging when listening in a non-native language [1], and gets even more challenging for both native and non-native listeners with increasing levels of noise (e.g., [2]).

The processes underlying the spoken-word recognition process are still largely unclear. However, spoken-word recognition theories agree on two central processes: all words that partly overlap with the input are activated simultaneously, and these words compete for recognition [3-6]. The number of activated words relates to word recognition times: an increase in the number of activated words results in an increase in competition, which in turn slows down word recognition [7].

There are several reasons why word recognition is harder in a non-native language than in one's native language. First, the phonemic repertoires of the native and non-native language differ (e.g., Dutch does not have the /æ/ as in English *marry*), resulting in inaccurate sound perception in non-native listening. This problem percolates upwards in the speech recognition process, leading to the spurious activation of additional words, not only in the non-native language (e.g., English *merry*) but also from the native language (e.g., Dutch *merrie*, E: *mare*) [8,9]. These spurious words are difficult to suppress, resulting in more competition [10], slowing down word recognition, and decreasing word recognition accuracy. Secondly, non-native vocabulary knowledge is obviously less rich or extensive than one's native vocabulary knowledge [1]. As a consequence, the spoken word might not even be available to the non-native listener. Listening in noise might magnify these problems for non-native listeners. Alternatively, it might be that native word selection strategies still hold (to some extent) during non-native word recognition in noise. These native word recognition strategies may help to restrain the increase in number of spuriously activated words when listening in noise in a non-native language.

One such native strategy is the use of both word-initial and word-final information for candidate word selection and word recognition [11-13].Word-initial information seems to be the most important of the two [3,6,13], at least in clear listening conditions. This could simply be due to the fact that word beginnings are heard before word endings and therefore receive most processing. Indeed, when listening in noise, word-final information seems to become relatively more important compared to clear listening conditions [13], while [12] suggested that word-initial and word-final information might even be equally important for native spoken-word recognition, in noise.

The central questions of this study are: 1) are non-native listeners able to use both word-initial and word-final information during spoken-word recognition; 2) if so, are these information sources equally important for word recognition when listening conditions deteriorate? Experience with the non-native language has been shown to affect word recognition in the non-native language, i.e., an increase in proficiency in the non-native language leads to an increase in recognition accuracy in that non-native language (e.g., [14]). Our third question is therefore whether proficiency in the non-native language is correlated to word recognition in noise and to the use of word-initial and word-final information in noise.

To address these questions, a group of native Dutch listeners participated in a word recognition experiment in which English words were presented in clean or, crucially, had their onsets or offsets masked by noise at three increasingly difficult signal-to-noise ratios (SNRs). When word onset was masked, word-initial information was no longer reliably available; when word offset was masked, word-final information was no longer reliably available. Recognition accuracies in the onset-masked and offset-masked conditions, in the three SNR conditions, and in clean were compared.

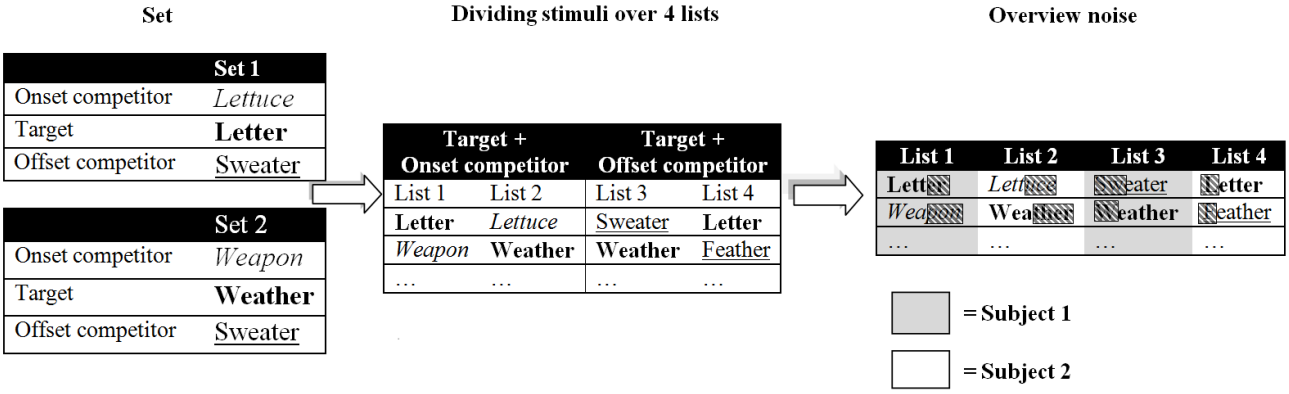

Figure 1: *The division of the stimuli over the lists. See text for information regarding this figure.*

2. Experimental set-up

2.1. Participants

Forty-seven participants, all native Dutch speakers, were drawn from the Radboud University Nijmegen participant pool, and were paid for their participation in the experiment. The mean age was 21.2 years (SD = 2.0; age range: 18 – 25; 39 females). The participants had an average of 8.1 years of formal English education (SD = 2.3; range: 6 – 16 years).

Proficiency of English of the participants was measured using the LexTALE task [15]. LexTALE is an un-speeded visual lexical decision task, measuring English vocabulary knowledge for medium to highly proficient speakers of English quickly and easily. The mean percentage correct was 65.1 (SD = 12.7; range: 40.7 – 96.6%).

2.2. Stimuli

The stimuli used in our word recognition task consisted of 42 sets of three partially overlapping English words, i.e., a target word (e.g., *letter*), an onset competitor, i.e., word-initial information was shared with the target word (e.g., *lettuce*), and an offset competitor, i.e., word-final information was shared with the target word (e.g., *sweater*); 126 words in total. Fourteen of these sets consisted of disyllabic words (e.g., *letter*; 42 words in total) and 28 sets consisted of monosyllabic words (e.g., *dog*; 84 words in total). Seven sets of disyllabic words were taken from the study of [6]. The other seven sets of disyllabic words were based on words from [16]. The sets of monosyllabic words were selected by searching through image databases [17], online rhyme dictionaries [18], and online scrabble dictionaries [19]. All words within a set had the same stress pattern according to Celex [20]. For the disyllabic target words, word frequencies ranged from 19 per million to 2166 per million. For the monosyllabic target words, word frequencies ranged from 24 per million to 13180 per million. Note that the here-presented experiment is part of a project that aims to shed light onto the importance of word-initial versus word-final information in different listening conditions by native and non-native listeners. The stimuli used in this study were chosen with this larger aim in mind.

The stimuli were produced by a male native English speaker, and recorded in a sound-attenuated booth at 44.1 kHz. Subsequently, the audio files were down sampled to 16 kHz to make them compatible with the noise file. Before noise was added to the stimuli, intensity of all word audio files was set at 60 dB SPL.

2.3. Adding noise

The amount of onset and offset masking was tailored to each target word and onset/offset competitor pair, such that the overlap between the target word and the onset/offset competitor remained unmasked. Take, for example, the target word *letter* and its onset competitor *lettuce*. For this pair [ɚ] of [lɛtɚ] and [əs] of [lɛtəs] are masked (referred to as the offset-masked condition), while for the target word *letter* and its offset competitor *sweater*, [l] and [sw] are masked, respectively (onset-masked condition). The shading of the letters in the right hand side of Figure 1 indicates the amount of masking for these three words (note that in the actual process phonemes were used to determine the deviation point between the two words, not letters). The mean overlap between target words and onset competitors was 2.45 phonemes. The mean overlap between target words and their offset competitors was 2.71 phonemes.

The noise that was used in this experiment was stationary speech-shaped noise (SSN). SSN is a pure energetic masker, i.e., parts of the speech signal are obscured and therefore less audible due to masking by the noise masker. It has a fixed spectrum and no significant temporal modulations [21]. The three SNRs that were used were -12 dB, -6 dB, and 0 dB. In order to add the noise to the stimulus files, boundaries were manually placed in the audio files at positive-going zero-crossings between the overlapping part of the word and the part that differed, e.g., lett|er - lett|uce for the offset-masked condition and l|etter – sw|eater for the onset-masked condition. Subsequently, an X is put in the tier to mark the part of the signal that has to be masked. The SSN noise is subsequently added to the sound files automatically using a PRAAT script [22]. The script places a random part of the noise signal on the marked part of the word. For the onset-masked stimuli, 200 ms of leading noise is added to the stimuli; for the offset-masked stimuli, 200 ms of trailing noise is added. A Hamming window is applied to the noise, with a fade in of 10 ms for onset masking and a 10 ms fade out for offset masking.

2.4. Lists

For the experiment, the words are divided over four separate "lists". Figure 1 shows the procedure we followed. The tables on the left hand side of Figure 1 represent the sets of words. The word in bold is the target word (e.g., **letter**), the word in italics is its onset competitor (*lettuce*), and the underlined word is its offset competitor (sweater). First, target words and their onset competitors are divided over the first two lists: the target word with the highest frequency is put in the first list,

the target word with the second highest frequency in the second list, the word with the third highest frequency in the first list, and so on. Next, the onset competitors are placed into these two lists, such that a target word and its onset competitor are never in the same list. So, lists 1 and 2 only contain target words and onset competitors (see also the left two columns of the middle part of Figure 1). Subsequently, the same target words and their offset competitors are divided over the third and the fourth list following the same procedure. Lists 3 and 4 thus only contain target words and offset competitors, as illustrated in the right two columns of the middle part of Figure 1. Finally, each list (42 stimuli) is divided into three blocks (14 stimuli) for the purpose of assigning a different SNR to each block to investigate the accuracy scores over the three listening conditions when the position of the noise is the same. The words were divided over the blocks on the basis of their word frequency to ensure that each block contained words with similar word frequencies.

2.5. Procedure

Participants were tested individually in a sound-treated booth. The stimuli were presented binaurally over closed headphones (Sennheiser HD 215 MKII DJ) at a comfortable sound level. The experiment consisted of three parts. One part only contained words with onset-masking (list 3 or 4; see the right of Figure 1); a second part only contained words with offset-masking (list 1 or 2). The order of these parts was counterbalanced across participants. The first blocks presented in the two parts got the same SNR (e.g., 0 dB), as do the second and third blocks of both parts (e.g., -6 dB for both second blocks and -12 dB for both third blocks). Part three, which always came last, consisted of all 84 words in the previous two parts without masking. The words within each block were randomized. After every block there was a self-paced pause. All odd-numbered participants got list 1 and 3 and all even-numbered participants got list 2 and 4 in counterbalanced order, (see the shading of the columns in the table on the right hand side of Figure 1). So, each participant only got targets from the offset-masked condition and competitors from the onset-masked condition, or targets from the onset-masked condition and competitors from the offset-masked condition. The task of the participant was to type in the word they thought they had heard. The experiment lasted approximately fifteen minutes. Afterwards, participants carried out the LexTALE task.

3. Results

All analyses were carried out using generalized linear mixed-effect models [23] containing both fixed and random effects, using the logit link function. Each analysis started with building the most complex model, i.e., a model containing all predictors and all possible interactions between the predictors in the fixed part of the model. Subsequently, interactions and predictors which did not reach the significance level (5%) were removed from the model one-by-one, starting with the least significant interaction or predictor. Each change in the fixed effect structure was evaluated in terms of model fit by means of a likelihood ratio test with the anova function in R. The dependent variable was correct versus incorrect recognition of items. Fixed factors were the SNR (-12 (on the intercept), -6, 0, clean) and the Position of Noise (onset-masked vs. offset-masked; the former is the reference category). Moreover, by-participants and by-stimuli random intercepts were added to the model. The model with the lower AIC value and, therefore, better model fit was retained. The best-fitting model only contains predictor variables and interactions that are significant. Here, only the final best-fitting models are reported.

First, homophones received the same orthographic transcription (e.g., *plain* and *plane*). Subsequently, the accuracy score for the clean condition was compared with the accuracy scores for the onset-masked and offset-masked conditions for SNR = 0. The results showed that participants recognized significantly more words correctly in the clean condition than in the onset-masked condition (β = 1.4173, *SE* = .1491, $p < .001$) and in the offset-masked condition (β = .8814, *SE* = 5.793, and $p < .001$).

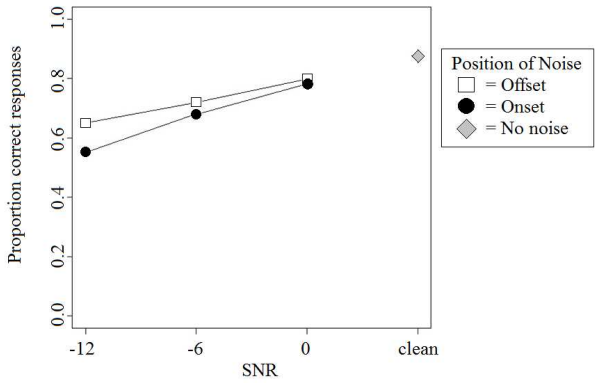

Figure 2: *Proportion of correct responses by position of noise and SNR.*

Table 1. *Fixed effect estimates for the best-fitting models of performance in the word recognition experiment.*

Fixed effects	***β***	***SE***	***p* <**
Model A: Position of Noise and SNR (n = 3948 observations)			
Intercept	0.6767	0.2464	.01
Position of Noise	-1.3072	0.2454	.001
SNR	0.4802	0.0812	.001
Position of Noise × SNR	0.2963	0.1121	.01
Model B: Proficiency (n = 3780 observations)			
Intercept	0. 7644	0.2522	.01
Position of Noise	-1.3859	0.2546	.001
SNR	0.4704	0.0847	.001
LexTALE	0.0140	0.0052	.01
Position of Noise × SNR	0.3072	0.1167	.01

3.1. Position of noise and listening conditions

To answer our research questions, accuracy scores for the two masking conditions and the three SNR conditions were analyzed. Figure 2 shows the proportion of correct responses for the different SNR conditions and the position of the noise (white squares for offset-masked, black circles for onset-masked, and the grey diamond for clean). As can be seen, when word onset was masked (bottom line) fewer words were recognized. Moreover, fewer words seem to be recognized with increasing difficulty of the listening conditions. Table 1 (see Model A) displays the parameter estimates in the best-fitting model of performance for the Position of Noise and the

SNR factors. In this model, the SNR is a covariate (continuous variable), since the covariate model had a better fit (AIC = 3563) than a model where SNR was included as a categorical variable with three values (AIC = 3567).

The statistical analysis showed an effect of Position of Noise (see Table 1): significantly fewer words were recognized when word onset was masked compared to when word offset was masked (onset: 67.1% correct, offset: 72.2%, averaged over all three SNR conditions). Moreover, an effect of SNR was found: more words were recognized with improving listening conditions (SNR = -12: 60.1% correct, SNR = -6: 69.9%, SNR = 0: 79.0%). Importantly, there is a significant interaction of SNR and Position of Noise, which means that significantly fewer words are recognized when the onset of the word is masked compared to when the offset is masked when listening conditions become harder.

3.2. Proficiency

Subsequently, the role of proficiency on the use of word-initial and word-final information in non-native word recognition in noise was investigated. Due to technical problems, the LexTALE scores of two participants were missing, their data were excluded from further analysis. First, the LexTALE scores were correlated with the accuracy scores for the three different SNRs for Position of Noise, separately. The scatterplots showed a positive relationship, although the only significant correlation (after Bonferroni correction) was found to be LexTALE with SNR = -12 dB with onset masking (r = .41, p < .01). This positive correlation means that the more proficient a non-native listener is, the more items are recognized when the onset of the word is masked.

A similar analysis to the analysis in the previous section was carried out, but with proficiency (normalized LexTALE score) as a third factor. The parameter estimates in the best-fitting model of performance are shown in Table 1 ("Model B"). No interaction effect between LexTALE and SNR was found, but there was a main effect of LexTALE: a higher LexTALE score led to an increase in the number of recognized words. The results for Position of Noise and SNR were similar to those found in Model A (see the beta values in Table 1). When LexTALE was included in the model, the model gained a better fit compared to the previous model (AIC = 3367 vs. AIC = 3563).

4. General discussion

The current study addresses the question whether non-native listeners use word-initial and word-final information like has been found for native listeners, and whether these information sources are equally important when listening conditions become harder. These questions are investigated through a word recognition experiment in English by Dutch listeners, where either a word's onset or offset is masked by speech-shaped noise in three different SNR conditions.

Native listeners have been shown to be able to exploit both word-initial and word-final information for the recognition of words [11-13]. Our results on a group of non-native listeners show that non-native listeners use both word-initial and word-final information when recognizing words in noise. Even in de hardest listening conditions (onset of the words masked with noise at an SNR of -12 dB), accuracy scores were on average above 50%. Non-native listeners thus are able to use information from anywhere in the speech signal for the recognition of words, i.e., if word onsets are (substantially) masked by noise, they are able to exploit information from the end of the word for recognition of the word and vice versa.

When listening conditions became harder, however, increasingly more errors were made when the onsets of the words were masked compared to when the offsets of the words were masked (see also the diverging lines in Figure 2). For correct recognition of a word, word-initial information thus seems to be more important than word-final information for Dutch non-native listeners of English with increasing difficulty of the listening conditions. This result seems to contradict the findings by [12]. In their study on native word recognition by Dutch listeners, where also word-onsets or word-offsets were masked with noise, they found that word-initial and word-final information were equally important. However, they only used one SNR, which they did not report (nor did they report the type of noise they used). Possibly, their SNR was similar to our SNR = 0 dB condition, where we also did not find a difference in importance of word-initial and word-final information. Their conclusion that "auditory word recognition is not more sensitive to word-initial than to word-final information" thus might not hold true when listening conditions deteriorate, at least not for non-native listeners.

At first sight, our results also seem to contradict those of [13], who found that word-final information becomes relatively more important when speech is less reliable. Whether this difference is due to differences in the point at which speech processing is investigated (online speech processing using eye-tracking vs. offline word recognition task) or to differences in listener population (native vs. non-native listeners) will be investigated in our next experiment.

In a second analysis, the role of proficiency in the non-native language on the use of word-initial and word-final information was investigated. Proficiency (as operationalized by LexTALE scores) only showed a main effect: more words were recognized with a higher proficiency, irrespective of whether the word was masked by onset or offset noise and irrespective of the SNR. This effect of proficiency is in line with other studies showing that an increase in proficiency leads to an increase in recognition accuracy in that non-native language (e.g., [14]). Importantly, the lack of an interaction of proficiency with Position in Noise shows that proficiency does not modulate the use of word-initial and word-final information in non-native word recognition.

To conclude, non-native listeners are able to exploit both word-initial and word-final information for the selection and recognition of words when part of the speech signal is masked. However, when listening conditions deteriorate, word-initial information is more important than word-final information, suggesting that the spoken-word recognition system is more sensitive to word-initial than word-final information during non-native listening when listening conditions become harder. This increasing effect occurred independently from the non-native language proficiency level of the participants, although proficiency level was correlated to test performance in general.

5. Acknowledgements

This research is sponsored by a Vidi-grant from the Netherlands Organisation for Scientific Research (NWO) to Odette Scharenborg. We thank Joop Kerkhoff and Ronald Fischer for their technical support. Also, we want to thank Polina Drozdova for her help in co-running this experiment and Alastair Smith for support in recording the stimuli.

6. References

[1] Garcia-Lecumberri, M.L., Cooke, M., Cutler, A., "Non-native speech perception in adverse conditions: A review", Speech Communication, 52(11-12):864-886, 2010.

[2] Cooke, M., Garcia Lecumberri, M.L., Barker, J., "The foreign language cocktail party problem: energetic and informational masking effects in non-native speech perception", Journal of the Acoustical Society of America, 123(1): 414–427, 2008.

[3] Zwitserlood, P., "The locus of the effects of sentential-semantic context in spoken-word processing", Cognition, 32:25-64, 1989.

[4] McQueen, J.M., Norris, D., Cutler, A., "Competition in spoken word recognition: Spotting words in other words", Journal of Experimental Psychology: Learning, Memory, and Cognition, 20(3):621-638, 1994.

[5] Gow, D.W., Gordon, P.C., "Lexical and prelexical influences on word segmentation: Evidence from priming", Journal of Experimental Psychology: Human Perception and Performance, 21(2): 344-359, 1995.

[6] Allopenna, P.D., Magnuson, J.S., Tanenhaus, M.K., "Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models", Journal of Memory and Language, 38:419-439, 1998.

[7] Norris, D., McQueen, J.M., Cutler, A., "Competition and segmentation in spoken word recognition", Journal of Experimental Psychology: Learning, Memory, and Cognition, 21(5):621-638, 1995.

[8] Weber, A., Cutler, A., "Lexical competition in non-native spoken-word recognition", Journal of Memory and Language, 50(1):1-25, 2004.

[9] Cutler, A., Weber, A. and Otake, T., "Asymmetric mapping from phonetic to lexical representations in second-language listening", Journal of Phonetics, 34(2):269-284, 2006.

[10] Broersma, M., Cutler, A., "Competition dynamics of second-language listening", Quarterly Journal of Experimental Psychology, 64:74-95, 2011.

[11] Slowiaczek, L.M., Nusbaum, H.C., Pisoni, D.B., "Phonological priming in auditory word recognition", Journal of Experimental Psychology: Learning, Memory, and Cognition, 13(1):64-75, 1987.

[12] Van der Vlugt, M.J., Nooteboom. S.G., "Auditory word recognition is not more sensitive to word-initial than to word-final stimulus information", Journal of the Acoustical Society of America, 81:41-49, 1986.

[13] McQueen, J.M., Huettig, F., "Changing only the probability that spoken words will be distorted changes how they are recognised", Journal of the Acoustical Society of America, 131(1):509-517, 2012.

[14] Imai, S., Walley, A.C., Flege, J.E., "Lexical frequency and neighborhood density effects on the recognition of native and Spanish-accented words by native English and Spanish listeners", Journal of the Acoustical Society of America, 117(2):896-907, 2005.

[15] Lemhöfer, K., Broersma, M, "Introducing LexTALE: A quick and valid Lexical", Behavior Research Methods, DOI: 10.3758/s13428-011-0146-0, 2011.

[16] Schock, J., Cortese, M.J., Khanna, M.M., Toppi, S., "Age of acquisition estimates for 3,000 disyllabic words", Behavior Research Methods, 44:971–977, 2012.

[17] Snodgrass, J.G. and Vanderwart, M., "A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity", Journal of Experimental Psychology: Learning, Memory, and Cognition, 6(2):174-215, 1980.

[18] Beeferman, D. "Rhymezone." Internet: http://www.rhymezone.com/.

[19] Shimoda, D. "Scrabblefinder." Internet: http://www.scrabblefinder.com/.

[20] Baayen, R. H., Piepenbrock, R., Van Rijn, H., "The CELEX lexical data base on CD-ROM". Philadephia, PA: Linguistic Data Consortium, 1993.

[21] Cooke, M., Garcia Lecumberri, M.L., Scharenborg, O., Van Dommelen, W.A., "Language-independent processing in speech perception: identification of English intervocalic consonants by speakers of eight European languages", Speech Communication, 52:954-967, 2010.

[22] Boersma, P., Weenink, D., "Praat. Doing phonetics by computer (Version 5.1)", 2005.

[23] Baayen, R.H., Davidson, D.J., Bates, D.M., "Mixed-effects modelling with crossed random effects for subjects and items", Journal of Memory and Language, 59(4):390-412, 2008.