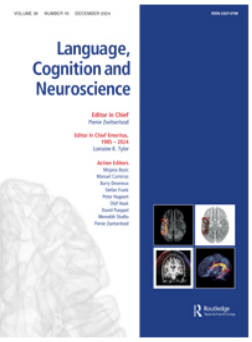

Routledge
Taylor & Francis Group

Language, Cognition and Neuroscience

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/plcp21

How lexical frequency, language dominance and noise affect listening effort – insights from pupillometry

Jens Schmidtke, Dana Bsharat-Maalouf, Tamar Degani & Hanin Karawani

To cite this article: Jens Schmidtke, Dana Bsharat-Maalouf, Tamar Degani & Hanin Karawani (20 Nov 2024): How lexical frequency, language dominance and noise affect listening effort – insights from pupillometry, Language, Cognition and Neuroscience, DOI: 10.1080/23273798.2024.2427832

To link to this article: https://doi.org/10.1080/23273798.2024.2427832

LANGUAGE, COGNITION AND NEUROSCIENCE
https://doi.org/10.1080/23273798.2024.2427832

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

How lexical frequency, language dominance and noise affect listening effort – insights from pupillometry

Jens Schmidtke [a,b], Dana Bsharat-Maalouf [c], Tamar Degani [c] and Hanin Karawani [c,d]

[a]Haifa Center for German and European Studies, University of Haifa, Haifa, Israel; [b]Herder Institut, Universität Leipzig, Leipzig, Germany; [c]Department of Communication Sciences and Disorders, University of Haifa, Haifa, Israel; [d]Cluster of Excellence Hearing4All, University of Oldenburg, Oldenburg, Germany

ABSTRACT
Acoustic, listener, and stimulus-related factors modulate speech-in-noise processes. This study examined how noise, listening experience, manipulated at two levels, native [L1] vs. second language [L2], and lexical frequency impact listening effort. Forty-seven participants, tested in their L1 Hebrew and L2 English, completed a word recognition test in quiet and noisy conditions while pupil size was recorded to assess listening effort. Results showed that listening in L2 was overall more effortful than in L1, with frequency effects modulated by language and noise. In L1, pupil responses to high and low frequency words were similar in both conditions. In L2, low frequency words elicited a larger pupil response, indicating greater effort, but this effect vanished in noise. A time-course analysis of the pupil response suggests that L1–L2 processing differences occur during lexical selection, indicating that L2 listeners may struggle to match acoustic-phonetic signals to long-term memory representations.

ARTICLE HISTORY
Received 21 June 2024
Accepted 31 October 2024

KEYWORDS
Spoken-word recognition; second language; bilingualism; listening effort; speech in noise

Introduction

Speech understanding is a complex cognitive process that requires transforming ambiguous acoustic signals into a hierarchy of representations, ranging from proper auditory processing of speech sounds to linguistic representations (Poeppel & Sun, 2021). Under favourable listening conditions, understanding the speech signal in real time as it unfolds happens seemingly effortlessly. However, speech understanding can become more taxing under less favourable listening conditions, such as when speech is masked by a competing signal like noise. In addition to these external, acoustic factors, speech understanding is influenced by listener internal factors such as language proficiency (Mattys et al., 2012; Pisoni, 2021). Moreover, stimulus-related factors such as lexical frequency contribute to speech understanding in that recognition of high frequency words is more robust to signal degradation compared to low frequency words (Mattys et al., 2012).

Here, we investigated the combined effects of acoustic, listener and stimulus-related factors on listening effort (Peelle, 2018). Specifically, we tested participants on single word recognition in their first and dominant language (L1) and in their less-dominant second language (L2) in quiet and noisy conditions. Critically, across these conditions, we examined how lexical frequency affected listening effort as indexed by pupil dilation, a method known to reflect listening effort (Zekveld et al., 2010).

Lexical frequency and listening in L2

Lexical frequency has been highlighted as a critical organising factor in the architecture of the mental lexicon, in that high frequency words are recognised faster and more accurately than low frequency words (Goldinger et al., 1989; Marslen-Wilson, 1987). This finding has been explained in terms of more entrenched representations in memory for high frequency items (Goldinger, 1998; also see Bybee, 1985, p. 117), connection strengths between sub-lexical and lexical levels of representation (Dahan et al., 2001), and a post-lexical decision bias (Luce & Pisoni, 1998). The frequency effect has also been studied as it relates to first and second languages (Duyck et al., 2008). Most research in this domain comes from

CONTACT Hanin Karawani hkarawani@staff.haifa.ac.il Department of Communication Sciences and Disorders, University of Haifa, Eshkol Tower, Office 815, 199 Aba Khoushy Ave, Haifa, 3498838, Israel

Supplemental data for this article can be accessed online at https://doi.org/10.1080/23273798.2024.2427832.

visual word recognition and speech production tasks, and a common finding across studies is that frequency effects are larger in the L2 (Gollan et al., 2008; Whitford & Titone, 2012).

This finding also seems to extend to spoken word recognition, although few studies exist. In one study, Shi (2014) tested English monolingual speakers and three groups of bilingual speakers with varying proficiency levels in English. Results showed that the correlation coefficient between perceptual accuracy and frequency was highest for the group with the lowest proficiency in English and lowest (and non-significant) for the monolingual group (also see Shi, 2015), suggesting that less proficient listeners are more affected by lexical frequency. Likewise, a pupillometry study that investigated listening effort during single word recognition in English found that lexical frequency had a larger effect on L2 listeners compared to L1 listeners (Schmidtke, 2014).

One way to explain the larger frequency effect in L2 is to focus on the accumulated experience listeners have with their languages. Specifically, in the case where the L2 was acquired later and/or is used less often compared to the L1, L2 words would be of lower frequency compared to L1 words (cf. Gollan et al., 2008). Due to the logarithmic property of the frequency effect (under this assumption, the difference between 1 and 10 occurrences per million words would be as dramatic as the difference between 10 and 100 occurrences per million), frequency effects should be more pronounced at the lower sections of the distribution, yielding stronger effects in L2 than in L1.

Lexical frequency and noise

The effect of background noise on listening effort is well established (e.g. Peelle, 2018). Critically, effects of lexical frequency have been documented in noise, such that in less favourable listening conditions high frequency words are recognised more accurately (Howes, 1957) and are recognised at lower speech-to-noise ratios (SNRs) compared to low frequency words (Pollack et al., 1959; Savin, 1963). Yet, relatively few studies have investigated the degree to which the frequency effect is modulated by noise. For example, Strauß et al. (2022), testing listeners in their L1, observed that reaction times decreased with increasing word frequency and with decreasing noise levels. However, no interactions were observed between the two factors.

Van Engen et al. (2020) investigated the effects of noise and word frequency in younger and older adults (in their L1) using the visual-world eye-tracking paradigm. They observed faster lexical access, as indexed by a higher probability of fixating the target picture, in quiet than in noise, and for higher lexical frequency. They also note an interaction between the two factors but the direction of this effect was not reported, so it is unclear how noise impacted the frequency effect.

Most relevant to the current study, Kuchinsky et al. (2023) used a dual-task paradigm to tap into listening effort, comparing recognition of low and high lexical frequency words under easier or more difficult SNRs. They further manipulated the difficulty of the secondary task. They found that lower lexical frequency was associated with overall slower responses to the secondary task, suggesting an effect of frequency on listening effort. Notably, when the secondary task was easier, noise and frequency interacted, with a larger frequency effect in more challenging listening conditions. However, when the secondary task was more difficult, these effects disappeared. This suggests that lexical frequency effects on listening effort are sensitive to task demands.

Lexical frequency, noise, and listening in L2

Thus far, the reviewed evidence suggests that the frequency effect is larger in L2 than in L1 (e.g. Duyck et al., 2008), and that it may be larger under more challenging listening conditions (Kuchinsky et al., 2023). Moreover, there is extensive evidence to suggest that the effect of noise is more pronounced in L2 compared to L1 (Bsharat-Maalouf & Karawani, 2022a, 2022b; Garcia Lecumberri et al., 2010; Scharenborg & van Os, 2019). This latter finding can be explained by exemplar models, which suggest that phonetic and lexical representations are built and strengthened through repeated exposure to specific instances or exemplars of words (Goldinger, 1998; Pierrehumbert, 2003). In the following we will refer to this as the (phonetic) *Entrenchment Hypothesis*. As individuals learn new L2 words, their representations will initially be weak, and they will behave like very rarely encountered L1 words. As learning progresses, representations of frequently encountered L2 words will strengthen and a frequency effect will emerge. Because L2 words have reduced frequency of experience relative L1 words (cf. Gollan et al., 2008), and that low frequency words are processed less efficiently in noise, the entrenchment hypothesis correctly predicts that word recognition in an L2 under noise conditions will result in higher listening effort compared to L1.

In this study, we examined how lexical frequency (a stimulus-related factor), noise (an acoustic factor), and language dominance (a listener-internal factor) jointly influence listening effort. Previous research has not directly explored the combined effects of these variables. For instance, Schmidtke (2016) tested monolingual and bilingual (L2) listeners in a speech-in-noise task, asking them to repeat the final word of sentences. The analysis revealed that accuracy was predicted by an interaction between word frequency and vocabulary size, with a larger difference between groups for low-frequency words. However, without a quiet condition, it remains unclear how noise interacts with these factors, and the focus on accuracy rather than listening effort limits the findings. Our study further expands on previous research by employing an online measure of language processing, providing finer detail on L1–L2 processing differences in noisy environments, as outlined in the next section.

Time course of lexical access

Exploring how stimulus-related, acoustic-related, and listener-related factors unfold over time represents a novel approach, as most studies examining the impact of noise on word recognition have focused on accuracy as the primary outcome. Additionally, our approach offers a new perspective by integrating insights from both the speech perception and spoken-word recognition literatures.

One common assumption is that spoken word recognition involves three basic mechanisms. (1) Incremental activation: Lexical items that partially match the unfolding speech signal become active. (2) Parallel activation and competition: Similar sounding words are activated in parallel according to the goodness of fit with the signal and prior experience of their occurrence. (3) Selection: Among the active candidate words, the one that best matches the signal and the context is selected (Kapnoula et al., 2024; Luce & Pisoni, 1998; Marslen-Wilson, 1991).

In recent years, the visual-world eye-tracking paradigm has been instrumental in showing that the process of recognising words from speech starts as soon as acoustic information becomes available. For example, Allopenna et al. (1998) showed that looks to the target picture started to increase 200 ms after target word onset, and given that it takes ~150 ms to launch a saccade, 200 ms is basically the earliest possible point for eye movements driven by the auditory input. In another study using the same paradigm, Dahan et al. (2001) manipulated the lexical frequency of target words and demonstrated that frequency effects were visible early on in a trial. This suggests that word frequency is not merely a result of a decision bias occurring post-lexical activation, but also plays a critical role during the initial stages of word recognition (also see Cleland et al., 2006).

These early frequency effects contrast with the predictions of the Neighbourhood Activation Model (NAM, Luce & Pisoni, 1998), which posits that frequency only acts as a response bias later on in the recognition process. However, at what point in time frequency effects emerge can also depend on task demands. For example, Winsler et al. (2018) observed early frequency effects in event related potential (ERP) data collected during an auditory lexical decision task, but when the task was semantic categorisation, frequency effects emerged later. Lastly, using an auditory lexical decision task, Dufour et al. (2013) found facilitative effects of lexical frequency on the P350 (before word offset) and the late N400 (after word offset). They posited that these two separate effects of frequency on the EEG signal can be mapped onto two different stages of word recognition. The early effect may relate to lexical activation and the later effect may relate to lexical selection.

In the present study, we used an online measure of processing, pupillometry, to investigate how word recognition is influenced by language dominance (L1 vs. L2), background noise, and lexical frequency.

Pupillometry

Like eye movements and EEG, pupillometry is an online measure of mental processes that has been used to investigate a range of different language-related tasks (Schmidtke, 2018; Schmidtke & Tobin, 2024). While the primary function of pupil constriction and dilation is to adjust focal distance and regulate the amount of light that enters the eye, pupillary dilation is also indicative of various cognitive processes (Strauch et al., 2022). Kahneman (1973) advanced the term *mental effort* to highlight the intensity dimension of attention, that is, how much attention is necessary to perform a task, and concluded that pupil dilations were an adequate measure of mental effort (p. 19).

One widely used application of pupillometry is in research on listening effort in noisy environments (Zekveld et al., 2018). For example, identifying words in noise is associated with larger pupil dilation than identifying nonspeech stimuli in noise, even when the difficulty of both tasks is matched (Kramer et al., 2013). Pupillometry has also been used to investigate L2 processing, and the general finding is that listening in L2 increases effort (for a recent review see Bsharat-Maalouf et al., 2023). Francis et al. (2018) recorded pupil dilations of Dutch speakers while they were listening to sentences in Dutch and English, their L2, with and without masking. Interestingly, there was no main effect of language on pupil dilation. Rather, listening appeared to be most effortful when the target language and the masking language were the same. Other studies found an effect, suggesting that listening in L2 is more effortful (Borghini & Hazan, 2018, 2020; Schmidtke, 2014). However, these studies employed a between-participant design and so individual differences in the pupil response may be responsible for the L2 effect (but see Bsharat-Maalouf et al., 2024).

Compared to the evidence for the effect of stimulus-related variables such as lexical frequency on processing coming from behavioural and neurological studies, relatively few studies have looked at the effect of these variables on the pupil response. The studies that exist suggest that the pupil response is sensitive, for example, to differences in frequency and neighbourhood density (Haro et al., 2017; McLaughlin et al., 2022; Schmidtke, 2014).

The current study

In the current study we tested the effects of lexical frequency, noise and language dominance on listening effort. To this end, Hebrew-English bilinguals listened to words in quiet and noise in their L1 (Hebrew) and L2 (English). Since these participants learned English in school with limited practice outside the classroom, they had less experience listening to English compared to Hebrew. In addition, the manipulation of word frequency allowed us to look at stimulus-specific effects of experience. High frequency words are by definition encountered more often than low-frequency words. Building on the theory that the more frequently words are encountered, the more robust their representations become in long-term memory, thus facilitating lexical access (Goldinger, 1998), we hypothesised that word recognition would be least effortful for high-frequency L1 words and most effortful for low-frequency L2 words. The entrenchment hypothesis mentioned above predicts noise and frequency to interact, so that noise affects low-frequency words more than high-frequency words, because the more robust a lexical representation is, the less susceptible it is to the detrimental effects of noise. This interaction would likely also be mediated by language: because processing efficiency of words in L1 is closer to ceiling, the interaction of frequency and noise is expected to be smaller in L1 than in L2.

Furthermore, we investigated whether these effects interact with time. A time course analysis can provide additional insight into processing differences between L1 and L2. Tracking the time course of the pupil response may indicate at which phase during spoken-word recognition processing lexical frequency effects emerge, and differences between L1 and L2 occur.

Methods

Participants

We tested 47 Hebrew (L1)-English (L2) university students (34 female, 13 male), with a mean age of 26.4 years (*SD* = 4.6). Participants grew up speaking Hebrew at home and learned English on average at age 7.4 (*SD* = 1.2) at school. Mean self-rated English proficiency was 7.4 (*SD* = 1.3) on a 0–10 scale. Participants reported being exposed to English 17.3% of the time (*SD* = 7.9). In addition, participants completed the *Multilingual Naming Test* (MINT; Garcia & Gollan, 2022) in Hebrew and in English, and the mean scores were 91.5% (*SD* = 5.0) and 62.8% (*SD* = 13.5), respectively.

Participant were pre-screened prior to coming to the testing session. Exclusion criteria were hearing loss, cataracts, cognitive or mental disorders, or a history of learning or language disabilities. In addition, participants could not participate if they were taking any pharmacological substances and were asked not to consume caffeine less than three hours before the experiment.

In terms of sample size, we aimed for a larger sample compared to previous pupillometry studies that investigated the frequency effect, for which the number was between 18 and 35 participants (see Haro et al., 2017; Kuchinke et al., 2007; Schmidtke, 2014). Further, following the recommendations of Brysbaert and Stevens (2018), we aimed to have at least 1600 observations per condition for each combination of language and noise. Frequency was examined as a continuous variable here.

Materials

Listening task

Stimuli. A total of 120 lexical items, all nouns, were chosen in each language that participants were expected to be familiar with (Hebrew and English). These were matched as closely as possible on different dimensions. Frequency information for Hebrew came from the OpenSubtitles 2018 corpus on SketchEngine.eu (Kilgarriff et al., 2014) and for English from SUBTLEX$_{US}$ (Brysbaert & New, 2009). The mean frequency per million for Hebrew was 44.9 (*SD* = 78.1, *Median* = 19.1, range = 0.1–490) and for English 57.3 (*SD* = 89.1, *Median* = 24.3, range = 0.8–554). A histogram of the log-frequency per million showed a normal distribution. The mean number of phonemes was 5.1 for Hebrew and 4.2 for English. Due to the different characteristics of each language, audio recordings for English words were slightly longer compared to the Hebrew recordings. Mean stimulus length was 881 ms ($SD = 115$) for English and 776 ms ($SD = 97$) for Hebrew words. To control for these differences, length (in ms) was entered as a control variable in all statistical analyses.

Out of the 120 lexical items in each language, each participant was presented with a subset of 80 items, 40 in quiet and 40 in noise, rotated so that all words were heard equally often across all participants. These single words were presented interspersed with 160 sentences in a pseudo-randomized order for a total of 240 stimuli (the results from the sentences will be reported in a separate study on sentence processing in noise[1]). The final word of each sentence was also taken from the same 120 lexical items so that for half of our experimental stimuli, the target word had been heard before in a sentence context. Because participants were asked to recall the whole sentence and the fact that the listening condition of the second presentation was always different from the first, we expected any carry-over effects from one block to the other to be minimal. This was confirmed by a mixed-model analysis of the accuracy data, which showed no main effect of presentation order or interaction with condition or language ($ps > .10$).

Recording. Stimuli were spoken by a native female speaker of each respective language, to avoid accented speech. Stimuli were recorded using JBL Tune 500BT headphones with a microphone, in a booth at 44.1 kHz and 32-bit. Speakers maintained a natural pace and neutral tone. Recordings were normalised for intensity through the root mean square function with Praat (Boersma & Weenink, 2014). Two native speakers per language confirmed the recordings' clarity and accuracy.

Noise manipulation. Speech-shaped noise was produced by filtering white noise to align with the average long-term spectrum of the stimuli across languages. The selection of a 0 dB signal-to-noise ratio (SNR) was informed by previous work indicating that such a level presents a manageable challenge across various participant profiles, including those who are multilingual (Bsharat-Maalouf & Karawani, 2022a; Cooke et al., 2010).

Proficiency test

The fast administration version of the Multilingual Naming Test (MINTSprint; Garcia & Gollan, 2022) was administered in Hebrew and in English following the standard procedure: Participants were shown 80 pictures of objects on a computer screen and were asked to name as many objects as they could in three minutes and correct responses were recorded. After this, participants were prompted to go through the pictures again and to name any that they did not name in the first round. The final score was calculated based on the percentage of correctly named objects from both rounds (for details, see Garcia & Gollan, 2022).

Procedure

The experimental session started with participants giving informed consent to participate according to the regulations of the university's ethics committee. Then they were seated in front of a computer screen in a recording booth with the experimenter present. They started with the listening task, either in L1 or L2, counterbalanced across participants, then completed language proficiency tests in the same language and then completed the listening task in the other language. At the end, they filled out a background questionnaire (Abbas et al., 2024).

Data were collected on the Eyelink Portable Duo (SR Research, Kanata, Ontario, Canada), monocularly from the pupil of the right eye at a sampling rate of 1,000 Hz. The camera was placed below the presentation screen and a chin rest was used to maintain the same distance to the camera across participants and to reduce head movement. The room was dimly lit, and the presentation computer screen maintained a constant grey background colour (RGB values: 225, 225, 225). During the experiment, the experimenter was present and monitored the eye-tracking data in real time on an experimenter screen.

The experiment started after a participant had read the instructions (given in Hebrew) and performed a nine-point calibration. Each trial started with 1000 ms of quiet or noise (depending on the listening condition), followed by the presentation of the stimulus, which was then followed by another 3000 ms of either quiet or noise. The fixation cross in the centre of the screen, which was present from the beginning of the trial until the end of the 3000 ms after stimulus offset, then turned into a question mark to prompt the participant to repeat what they had heard. Verbal responses were recorded on a recording device and coded for accuracy offline by the second author. No feedback was given to participants, and breaks were allowed upon request.

Preprocessing of pupil data

Data of all participants were aggregated and pre-processed using Eyelink's Data Viewer. Blinks were automatically removed and data in the 100 ms preceding and trailing a blink event were set to missing values. Data were then downsampled from 1000 Hz to 100 Hz, which corresponds to a temporal resolution of 10 ms, and exported to a CSV file. Further processing was done in R (R Core Team, 2024). Trials with more than 25% missing data were excluded (0.2% of all trials) and missing values were replaced by linear interpolation. Lastly, a four-point moving average smoothing filter was applied.

The resulting data set was then submitted to an outlier analysis. First, the inter-quartile range (IQR) was calculated for each participant and the lower and upper bounds were defined as the first and the third quartile, respectively, ± 2*IQR. Observations that were outside these bounds were marked as outliers. Per participant, the percentage of outliers ranged between 0.1% and 3.7%. For the four combinations of language and listening condition, the percentage of outliers ranged between 0.9% and 1.6%. Trials with more than 10% outliers were excluded from the analysis (4.2% of all trials). Excluding outliers did not change the pattern of results but improved model fits as suggested by a visual inspection of scatter plots of model residuals against predicted values (i.e. points were randomly distributed).

Analysis

The accuracy data were analysed in JASP (2024) using a Bayesian generalised mixed model because the analysis with lme4 resulted in a singular fit due to the lack of variance between participants (accuracy was generally high). The model converged using the default settings.

All remaining analyses were conducted on baseline corrected pupil measurements obtained from the Eyelink eye tracker. These measurements are reported in arbitrary units, as they correspond to the number of pixels occupied by the pupil, which can vary depending on factors like camera settings and distance. The baseline was calculated for each trial individually by taking the mean pupil size of the 200 ms interval preceding stimulus presentation. This value was then subtracted from all subsequent measurements in a trial. Thus, the dependent variable we used is the task-evoked pupillary response (short: pupil response).

Analyses were carried out using R’s lme4 package (Bates et al., 2015) and Python’s statsmodels library (Seabold & Perktold, 2010). Additional R packages used included car, dplyr, ggplot2, and sjPlot. Linear mixed effects models were fit to the pupil data. Fixed effects were language (L1: Hebrew/ L2: English), listening condition (quiet/noise) and word frequency (the log of frequency per million), and all interactions. Recognition accuracy (correct/incorrect) and stimulus length (in ms) were entered as control variables. Categorical predictors were sum-coded (−0.5/0.5) so that the estimate shows the mean of the variable, and continuous predictors were standardised so that the estimate is the change in the outcome variable associated with a change of 1 *SD* in the predictor variable. We included random intercepts for participants and random slopes for language and listening condition within participants. Random intercepts for items were not included because such models did not converge, likely because word frequency and stimulus length as item-level predictors already captured much of the item variance.

Different methodologies exist for analysing pupil data, each offering varied insights into the underlying cognitive mechanisms (Książek et al., 2021). For instance, when the primary focus of a study is on the listening effort associated with different SNRs, calculating the mean pupil dilation over the entire trial duration is often sufficient to detect significant effects. In this study, however, our interest lay on the time course of effects. Specifically, we aimed to identify significant changes in the pupil response linked to our predictors throughout the duration of a trial. Given our extensive dataset and the complexity of our model (i.e. the nested structure), the use of generalised additive models, as recommended by van Rij et al. (2019), proved too computationally demanding. Consequently, we employed a cross-validation method outlined in Mathôt and Vilotijević (2023), utilising the Python library Time Series Test (v. 0.12.0). For a more detailed methodology, we refer readers to the library documentation (https://github.com/smathot/time_series_test).

The cross-validation time series test works as follows: The dataset is segmented by assigning each trial to one of four folds in an interleaved manner (using the default settings), with each fold encompassing 25% of the data. Then, four separate linear mixed-effects models are conducted on each time bin across the training set comprising 75% of the data, systematically excluding one fold for the test set in each model. The sample (i.e. the bin) for which the strongest effect is observed in the training set is then taken from the test set of each fold to construct a new dependent variable, aggregating these peak effects across all test sets. The final step involves running a linear mixed-effects model on this aggregated dependent variable to evaluate the consistency and robustness of the observed effects. This procedure is iteratively applied to each main effect and interaction term in the model.

This cross-validation approach is designed to control for multiple comparisons by using an independent test set to verify the effects identified in the training set. By focusing only on effects that are consistently observed across multiple, independent folds, we reduce the likelihood of Type I errors, ensuring that reported effects are robust and not due to random noise.

The output of the time series test highlights the samples that were tested in each of the four models (i.e. where the strongest effect was detected in each subset) and presents the results from the mixed-effects model analysis on these consolidated samples. As a supplementary analysis, we conducted mixed-effects models as specified above in R on each 10 ms time bin across the entire trial, resulting in a total of 350 models. This analysis provides intervals during which an effect is significant, however, it tends to yield more spurious results compared to the cross-validation method. Therefore, we only report intervals where significant effects coincide with those identified through cross-validation and spanned at least 200 ms, thus ensuring greater reliability of the findings.

Results

Descriptive statistics for perceptual accuracy are shown in Table 1 and the model output is shown in Table A1. A mixed model run on the accuracy data showed that only the main effect of listening condition was significant, *b* = 2.83, 95% CI = [2.01, 3.82], indicating that participants were less accurate in the noise condition than in the quiet condition regardless of the tested language and the lexical frequency of the stimulus.

For the pupil data, the effect of the experimental predictors over time is visualised in Figure 1. To ensure that any effects we observed were driven by the stimuli and not any confounding variables, we first analysed the mean pupil response in the first 200 ms of all trials. In this time window, there should be no effects driven by the stimulus, given that there is a delay between stimulus onset and the pupil response.[2] The model results confirmed that none of the effects related to stimulus (language or frequency) were significant (see Table A2 for the model output).

The results of the cross-validation analysis and the subsequent time-bin analysis are reported in Table 2. The three-way interaction between language, condition, and frequency was significant from 1260 ms to 2050 ms. The interaction between condition and language was significant from 2260 ms to 3500 ms, and Figure 1 suggests that this interaction was driven by a more sustained pupil response in the noise condition when the language was English (L2). That is, compared to the Hebrew and the quiet conditions, the pupil remained dilated relatively longer. Moreover, the interaction between language and frequency was significant from 1380 ms to 2830 ms. The main effect of language shows that the pupil response to English (L2) words was generally greater than to Hebrew (L1) words starting from 780 ms till 3500 ms. Noise resulted in overall larger responses compared to quiet. Here, we observed two windows in which the effect was significant: An early window from 140 ms to 410 ms and a later window from 890 ms to 3290 ms (Figure A1 shows the *p*-value of all effects for each 10 ms time bin).

Table 1. Accuracy data for word repetition.

		Hebrew (L1)		English (L2)	
		Mean	95% CI	Mean	95% CI
Quiet	High frequency	99.8%	[99.4, 100]	99.8%	[99.5, 100]
	Low frequency	99.9%	[99.7, 100]	99.8%	[99.4, 100]
Noise	High frequency	94.3%	[92.8, 95.8]	93.2%	[91.6, 94.7]
	Low frequency	94.8%	[93.4, 96.2]	92.1%	[90.2, 93.8]

Note: Values in square brackets represent the lower and upper bounds of the bootstrapped 95% confidence interval around the mean. Frequency was used as a continuous variable in the statistical analyses. Low and high frequency in this table is based on a median split.

To further investigate the locus of the three-way interaction, we ran a model with the mean pupil response during the interval during which it was significant (1260 ms–2050 ms) as the dependent variable (Table A3). Subsequent pairwise comparisons suggested that the frequency effect was only present when the language was *English (L2)* and the condition *Quiet* (Table A4). We also further investigated the language by frequency interaction and again reduced the data to the mean response during the interval during which it was significant (1380 ms–2830 ms). The results showed a significant interaction between language and frequency in English, *b* = 1.44, 95% CI = [0.16, 2.72], *p* = .027 (see Table A5), but no effect in Hebrew (Table A6).

Discussion

Previous research suggests that listening in L2 is more effortful compared to listening in L1, especially when it comes to listening in noise (Bsharat-Maalouf et al., 2023). The purpose of the present study was to further investigate the cause of these differences. The results suggest that accumulated language experience is a contributing factor to listening effort in L2. The recognition of L2 words that are relatively infrequent, based on a subtitle corpus, was associated with a larger pupil response and thus greater listening effort compared to more frequent words. Noise interacted with frequency, indicating that high frequency only facilitated recognition in quiet conditions but not in noise. In L1, however, there was no evidence that frequency influenced listening effort, whether in noise or in quiet. In the following sections, we first discuss the frequency effect in L2 and then explore possible reasons for its absence in L1. In a final section, we discuss what additional insights can be gained from analysing the time course of effects, as done in this paper, compared to a more traditional peak-picking analysis.

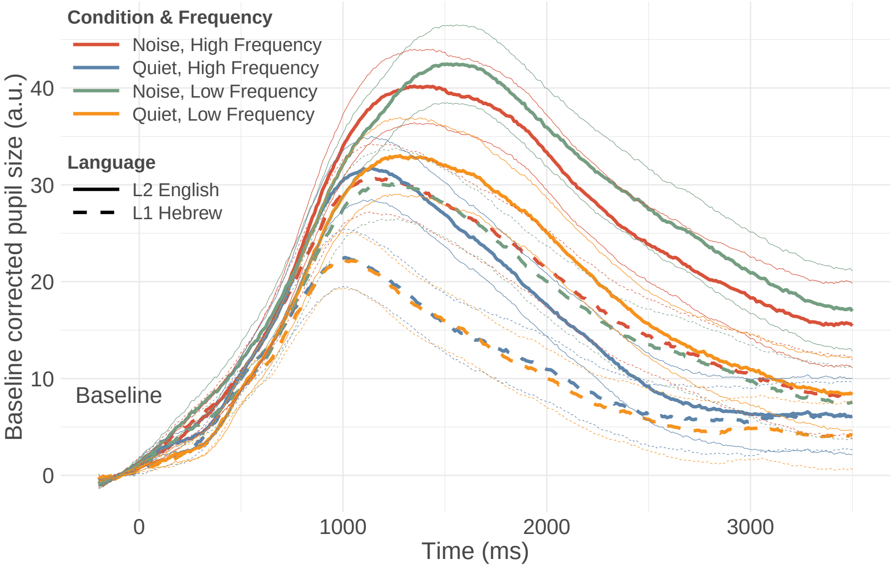

Figure 1. Baseline corrected pupil response to high and low frequency words presented in quiet and in noise in L1 Hebrew and L2 English.

Note: Baseline-corrected pupil response as a function of time averaged over all trials by listening condition, language (Hebrew L1, English L2), and lexical frequency. Frequency was divided into high and low based on a median split. The shaded areas represent the Standard Error of the Mean. A.u. = arbitrary units.

Word frequency effects in L2

The observation of a larger frequency effect in L2 processing as compared to L1 aligns with theories that attribute these differences to varying levels of language experience (Diependaele et al., 2013; Gollan et al., 2008; Schmidtke, 2016). These theories suggest that the cognitive mechanisms underlying language processing are fundamentally similar across L1 and L2, but that the extent of exposure and familiarity with the language modulates these processes, making frequency effects more pronounced in L2.

Table 2. Results of the cross-validation and bin analysis.

Cross-validation analysis				Time bin analysis
Effect	z	p	Peak effect (ms)	Interval (ms) with $p < .05$
Condition (quiet/ noise)	2.21	.027	1550, 1490, 1730	[140, 410], [890, 3290]
Language (He/En)	−2.58	.010	1640, 1650	[780, 3500]
Word frequency (WF)	0.16	.872	1640, 790	–
Condition × language	−3.14	.002	3050, 3140	[2260, 3500]
Condition × WF	1.53	.126	1920, 1620, 2910, 1170	–
Language × WF	2.67	.008	1760, 2030, 1890	[1380, 2830]
Condition × language × WF	−2.05	.041	1860, 1400, 1570	[1260, 2050]
Accuracy	6.04	<.001	2420	[1310, 3500]
Length	−0.12	.908	110, 100, 3390, 1200	–

Note: Accuracy and length were entered as control variables (and shown in grey). For the time-bin analysis, intervals that span more than 200 ms are reported when the effect was also significant in the cross-validation analysis (see Analysis section). WF = word frequency, He = Hebrew (L1), En = English (L2).

The present study adds to the literature by investigating the combined effect of word frequency and noise. If listening in noise is more effortful in L2 because of less experience compared to the L1, then we may expect that this effect would also be observed within a language as between frequently and less frequently encountered words. We found an interaction; however, the direction of this interaction was unexpected. Although the combined effects of word frequency, noise and language (L1 vs. L2) have not been investigated before, it is known that recognition accuracy for low frequency words decreases as the SNR decreases (Howes, 1957). Based on this finding, the expectation was that noise would make the recognition of low frequency words more effortful whereas noise would have a less detrimental effect on high frequency words. Instead in the current study, in noise, there was no advantage for high frequency words over low frequency words.

A possible tentative explanation for this finding comes from a word learning study by Creel and colleagues (Creel et al., 2012). In their study, participants learned new words that were presented either in noise or in quiet. The researchers found that recognition accuracy for these newly learned words was highest when the testing phase conditions matched those of the exposure phase: Words learned in noise were better recognised in noise, and the same was true for words learned in quiet. This suggests that participants did not encode abstract representations of the memorised words but rather formed context-specific representations that were tied to the conditions under which the words were learned. For our study, this implies that listeners may need exposure to noisy listening conditions for a frequency effect to emerge in these conditions. For example, our participants indicated that most of their exposure to English, their L2, came from media consumption, whereas exposure in more natural contexts, such as work or conversations with friends and family, was limited.[3] Thus, it is possible that a minimum amount of experience with noisy L2 stimuli is needed in order to allow an effect to be seen in such conditions. Once such a threshold is reached, we may expect to observe frequency effects in noise. Future studies with more variable language experience profile would shed light on this issue.

Absence of a frequency effect in L1

While the extant literature on the word frequency effect suggests that the effect is generally smaller in L1 compared to L2, the absence of the effect in the present data set was unexpected (though see Mor & Prior, 2020, who found a frequency effect in L2 English but not in L1 Hebrew on lexical decision times). Previous pupillometry studies that tested L1 speakers found the effect (see Haro et al., 2017; Kuchinke et al., 2007; Schmidtke, 2014). One possible explanation for our finding comes from Brysbaert et al. (2018). They show that the shape of the frequency effect is dependent on participants' vocabulary knowledge (ibid, Figure 2, p. 48). Vocabulary size not only influenced the size of the frequency effect but also the frequency range in which it was observed most strongly. Given that our participants had a large vocabulary in L1 as indicated by their MINT score, it may be necessary to include more items from the low end of the frequency scale to capture the effect. In other words, the null effect may reflect a ceiling effect that stems from the participants' overall high familiarity with the tested stimuli in the L1. Future studies in which lower frequency L1 words are sampled may reveal frequency effects on listening effort even in the L1.

The absence of the effect in L1 could also be the result of uncontrolled characteristics of the stimuli in Hebrew (L1). To rule out this possibility, we re-analysed previous data (see Table A7 and Figure 2 for a visual display of the data) collected from a group of Arabic native speakers with L2 Hebrew who were tested with the same Hebrew stimuli as in the present study (for a detailed description of the participants see Bsharat-Maalouf et al., 2024). The results of the analysis of this second data set align with the present study in that there was a frequency effect when Hebrew was the listeners' L2. This suggests that there is nothing inherent to the Hebrew stimuli which precludes a frequency effect. Instead, the absence of the effect in the L1 Hebrew listeners in the current study may indeed be due to a ceiling effect.

In conclusion, the results of the present study suggest that word frequency is a significant factor contributing to increased listening effort in L2. This finding underscores the crucial role of listeners' accumulated experience and familiarity with words, reflecting how repeated exposure and the frequency of encountering specific words can shape and refine lexical representations over time.

The time course of effects

In the present study, we opted for a different kind of analysis compared to most previous pupillometry studies of language processing. Traditionally, peak amplitude and peak latency of the pupil response are extracted on a trial basis which are then analysed (e.g. Zekveld et al., 2010). These measures have been shown to be sensitive to a range of different manipulations and suffice for many purposes. However, comparing different ways to analyse pupil data, Książek et al. (2021) suggested that these measures "provide less insight into the underlying processes of the pupil response (e.g. as a function of time)" (p. 13). On the other hand, cluster-based permutation tests, which are similar to the cross-validation test used in the present analysis, provide "more detailed information […] on the temporal profiles of the [task elicited pupil response]" (p. 14) (also see Hershman et al., 2023). Einhäuser (2017) even concludes that the pupil response "presents a time-varying signal that seems almost as rich as event-related potentials in EEG" (p. 163). Based on our analysis, we can make important observation as to how the pupil response relates to lexical access (cf. Rojas et al., 2024). To our knowledge, this has not been done before and can be valuable for future research to be able to make more specific predictions about the timing of effects.

As mentioned in the introduction, ERP studies of spoken word recognition found frequency effects in early and late components of the electrophysiological signal (Dufour et al., 2013; Winsler et al., 2018). In contrast, in the present study, frequency effects only occurred after word offset (cf. Table 2). This suggests that the pupil response during spoken word recognition represents later processes, such as lexical selection, rather than early stages of lexical access (activation and competition, Marslen-Wilson, 1987). It is true that

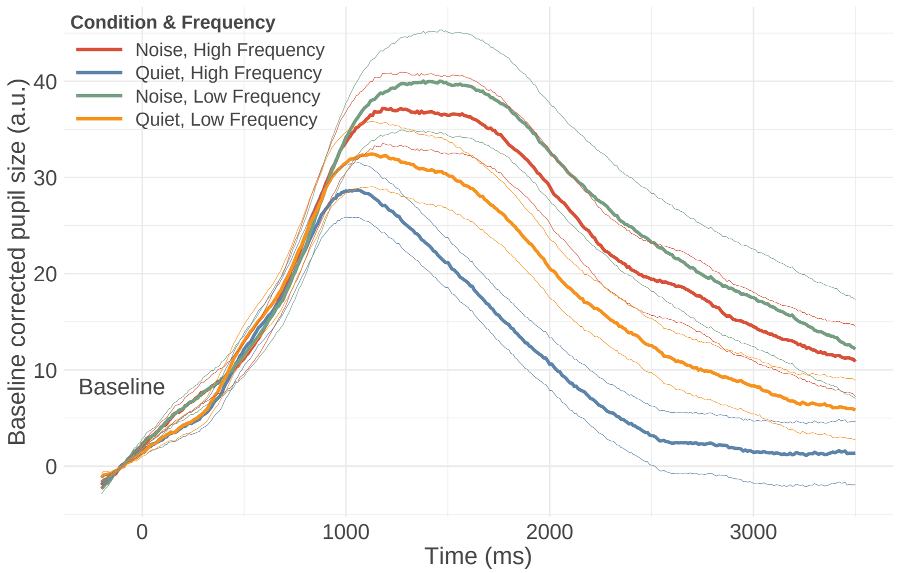

Figure 2. Baseline corrected pupil response to high and low frequency words presented in quiet in noise for the Arabic-Hebrew speakers in their L2 Hebrew.

Note: Baseline-corrected pupil response as a function of time averaged over all trials by listening condition and lexical frequency. Frequency was divided into high and low based on a median split. Participants were Arabic native speakers tested in their L2 Hebrew. The shaded areas represent the Standard Error of the Mean.

the pupil response is rather slow, certainly compared to EEG, but a stimulus-evoked pupil response can be observed as early as ~320 ms after stimulus onset (Hoeks & Levelt, 1993). Consequently, the fact that we found a "late" interaction between language and frequency may allow us to pinpoint the locus of processing differences between L1 and L2 word recognition. If the frequency effect observed here is indicative of lexical selection (Connine et al., 1993; Luce & Pisoni, 1998) – the process by which the correct word is selected from activated candidates – then the difficulty in L2 listening may be due to greater uncertainty as to which word was heard from among similar sounding candidates. This, in turn, may require more explicit processing. Next, we discuss how three models of language processing may account for our interpretation.

The Ease of Language Understanding (ELU) model (Rönnberg et al., 2022) asserts that listening effort arises from resolving mismatches between the signal and internal representations. While speech understanding is assumed to be effortless under ideal listening conditions, mismatches caused by noise or other factors require more explicit processing, thus demanding greater cognitive resources. If less frequent words have less precise representations in memory – especially in L2 contexts – then the processing of low-frequency words will also be associated with greater effort.

The computational model of bilingual word recognition, *Multilink* (Dijkstra et al., 2019; Dijkstra et al., 2023), represents lexical frequency effects by assigning lower resting activation levels to low-frequency items compared to high-frequency ones. To reflect the generally lower exposure of (unbalanced) bilingual speakers to L2 items versus L1 items, the resting activation levels of L2 items are adjusted downward relative to L1 items (in Multilink, this is achieved by dividing resting activation levels based on corpus frequencies by a constant factor of 4). As a result, L2 items typically require more cycles to reach the recognition threshold compared to L1 items with equivalent corpus frequencies, and the frequency effect is already observable during early cycles. Thus, this mechanism does not account for the late frequency effects in our study. One possible explanation is that pupil dilation, unlike eye-movements in the visual-world paradigm (Tanenhaus et al., 2000), does not directly represent lexical activation but rather reflects differences in uncertainty about the heard word, as suggested in the previous section. Therefore, a possible metric that aligns with pupil dilation may be the ratio of target word activation to all other active candidates. A higher ratio would indicate less uncertainty, which may translate into less listening effort.

The concept of uncertainty is effectively captured in Bayesian models of word recognition (Norris et al., 2016). Bayesian inference posits that the probability of recognising a specific word from the acoustic signal depends on two factors: the word's prior probability (its frequency) and the likelihood of the signal if that word was spoken. This framework accounts for variations in our study – namely, noise, word frequency, and language dominance.

Noise introduces uncertainty by masking phonetic cues, while low-frequency words increase uncertainty due to listeners' limited experience with them. This effect is even more pronounced for L2 listeners, who generally have less exposure to their second language, leading to greater uncertainty in recognising words and interpreting masked speech. Overall, Bayesian models may provide a compelling explanation for listening effort, as this effort arises from uncertainty in processing the acoustic signal (for studies linking response uncertainty to pupil dilation see Friedman et al., 1973; Richer & Beatty, 1987).

Limitations and future research

While the present study offers valuable insights, there are areas that future research can build upon. One consideration is the relatively small effect size observed for the frequency effect (compared to the effects of noise and language). A post-hoc power analysis using resampling with replacement ($n = 1000$) indicated that the three-way interaction between language, condition, and frequency (reported in Table A2) was replicated in 61% of cases. To enhance the robustness of these findings, future studies could benefit from including a greater number of stimuli in the low to medium frequency range, where the frequency effect is anticipated to be strongest (Brysbaert et al., 2018) even in the L1. This approach would not only target the expected effect more precisely but increasing the total number of stimuli would also improve power (Brysbaert & Stevens, 2018).

A different kind of masker could be used to establish whether the detrimental effects of noise and frequency observed here, generalise to multi-talker babble noise or informational masking. The task could be made more challenging by using a lower SNR. However, a more difficult task creates new problems because the researcher must decide how to deal with inaccurate trials. When a task is perceived as too difficult, participants may disengage (Herrmann & Johnsrude, 2020) and so pupil dilation on those trials would no longer be indicative of cognitive processes associated with lexical access. On the other hand, by excluding inaccurate trials, one would probably exclude a disproportionate number of low frequency items (since these tend to be more difficult to perceive) and so lose valuable variance in the data. In the present study we focused on single words because in a sentence context, the pupil response to the target word would also be influenced by the preceding words. However, it is important to test whether the present results generalise to more natural listening situations such as sentence or even passage comprehension, although controlling for possible confounding variables would be challenging.

Conclusion

This study showed that listening in L2 is more effortful compared to L1 and more effortful in noise than in quiet, even when recognition accuracy is relatively high, confirming previous studies (Bsharat-Maalouf et al., 2023). The purpose of this investigation was to examine whether listening in noise in L2 is harder because of less experience with the L2. The results showed that word frequency contributed to listening effort in L2, evidenced by a negative slope when regressing word frequency on the pupil response, indicating that lower word frequencies were associated with increased pupil dilation in the L2, at least under optimal listening conditions. For words presented in noise, we did not observe a processing advantage for high frequency words, a finding that should be investigated further in future studies. The timing of the frequency effect suggests that L1–L2 differences may arise during lexical selection, likely due to greater uncertainty about the heard word.

Notes

1. The full list of stimuli can be found at https://osf.io/zkvu3.
2. We do not assert that a 'cognitive' pupil response invariably begins at exactly 200ms after stimulus onset; rather, we assume that this duration represents the minimum time required for such a response to occur (see Mathôt & Vilotijević, 2023).
3. In an exploratory analysis we investigated this hypothesis further by examining whether individual differences in self-reported exposure to English or the English MINT vocabulary score would predict the size of the frequency effect in English. However, these variables did not explain additional variance when entered into the model reported in Table A4, potentially due to reduced variability on these dimensions in the current sample.

Acknowledgements

The data presented in this study were initially presented at the 33rd Annual Conference of the European Second Language Association (EuroSLA), held in Montpellier, France in 2024.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This study was supported by the Israel Science Foundation (ISF) [grant number: 2031/22 awarded to HK] and a fellowship by the Alexander von Humboldt Foundation (awarded to HK).

Ethic approval statement

The funding organizations had no role in the design and conduct of the study; in the collection, analysis, and interpretation of the data; or in the decision to submit the article for publication; or in the preparation, review, or approval of the article. Ethical IRB approval was obtained from the ethical committee of the University of Haifa [#243/18].

Data availability statement

A list of the stimuli along with data to reproduce Figures 1 and 2 can be found here: https://osf.io/n8rhg/.

ORCID

Jens Schmidtke http://orcid.org/0000-0001-7428-3478
Dana Bsharat-Maalouf http://orcid.org/0000-0002-6561-3364
Tamar Degani http://orcid.org/0000-0003-2604-8523
Hanin Karawani http://orcid.org/0000-0003-1346-8502

References

Abbas, N., Degani, T., Elias, M., Prior, A., & Silawi, R. (2024). Multilingual language background questionnaire in Hebrew. *PsyArXiv*. https://osf.io/preprints/psyarxiv/jfk8b

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*(4), 419–439. https://doi.org/10.1006/jmla.1997.2558

Bates, D. M., Maechler, M., Bolker, B., & Walker, S. (2015). Lme4: Linear mixed-effects models using S4 classes. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Boersma, P., & Weenink, D. (2014). *Praat: Doing phonetics by computer* (Version 5.3.51) [Computer program] (5.3.51). praat.org

Borghini, G., & Hazan, V. (2018). Listening effort during sentence processing is increased for non-native listeners: A pupillometry study. *Frontiers in Neuroscience*, *12*(152), 1–13. https://doi.org/10.3389/fnins.2018.00152

Borghini, G., & Hazan, V. (2020). Effects of acoustic and semantic cues on listening effort during native and non-native speech perception. *The Journal of the Acoustical Society of America*, *147*(6), 3783–3794. https://doi.org/10.1121/10.0001126

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science*, *27*(1), 45–50. https://doi.org/10.1177/0963721417727521

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*(1). https://doi.org/10.5334/joc.10

Bsharat-Maalouf, D., Degani, T., & Karawani, H. (2023). The involvement of listening effort in explaining bilingual listening under adverse listening conditions. *Trends in Hearing*, *27*. https://doi.org/10.1177/23312165231205107

Bsharat-Maalouf, D., & Karawani, H. (2022a). Learning and bilingualism in challenging listening conditions: How challenging can it be? *Cognition*, *222*(2022), 105018. https://doi.org/10.1016/j.cognition.2022.105018

Bsharat-Maalouf, D., & Karawani, H. (2022b). Bilinguals' speech perception in noise: Perceptual and neural associations. *PLOS ONE*, *17*(2), e0264282. http://dx.doi.org/10.1371/journal.pone.0264282

Bsharat-Maalouf, D., Schmidtke, J., Degani, T., & Karawani, H. (2024). Through the pupils' lens: Multilingual effort in first and second language listening. *Ear and Hearing*. https://doi.org/10.1097/AUD.0000000000001602

Bybee, J. L. (1985). *Morphology* (Vol. 9). John Benjamins. https://doi.org/10.1075/tsl.9

Cleland, A. A., Gaskell, M. G., Quinlan, P. T., & Tamminen, J. (2006). Frequency effects in spoken and visual word recognition: Evidence from dual-task methodologies. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(1), 104–119. https://doi.org/10.1037/0096-1523.32.1.104

Connine, C., Titone, D., & Wang, J. (1993). Auditory word recognition: Extrinsic and intrinsic effects of word frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(1), 81–94. https://doi.org/10.1037/0278-7393.19.1.81

Cooke, M., Garcia Lecumberri, M. L., Scharenborg, O., & van Dommelen, W. A. (2010). Language-independent processing in speech perception: Identification of English intervocalic consonants by speakers of eight European languages. *Speech Communication*, *52*(11–12), 954–967. https://doi.org/10.1016/j.specom.2010.04.004

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2012). Word learning under adverse listening conditions: Context-specific recognition. *Language and Cognitive Processes*, *27*(7–8), 1021–1038. https://doi.org/10.1080/01690965.2011.610597

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology*, *42*(4), 317–367. https://doi.org/10.1006/cogp.2001.0750

Diependaele, K., Lemhöfer, K., & Brysbaert, M. (2013). The word frequency effect in first- and second-language word recognition: A lexical entrenchment account. *Quarterly Journal of Experimental Psychology*, *66*(5), 843–863. https://doi.org/10.1080/17470218.2012.720994

Dijkstra, T., Peeters, D., Hieselaar, W., & Van Geffen, A. (2023). Orthographic and semantic priming effects in neighbour cognates: Experiments and simulations. *Bilingualism: Language and Cognition*, *26*(2), 371–383. https://doi.org/10.1017/S1366728922000591

Dijkstra, T., Wahl, A., Buytenhuijs, F., Van Halem, N., Al-Jibouri, Z., De Korte, M., & Rekké, S. (2019). Multilink: A computational model for bilingual word recognition and word translation. *Bilingualism: Language and Cognition*, *22*(4), 657–679. https://doi.org/10.1017/S1366728918000287

Dufour, S., Brunellière, A., & Frauenfelder, U. H. (2013). Tracking the time course of word-frequency effects in auditory word recognition with event-related potentials. *Cognitive Science*, *37*(3), 489–507. https://doi.org/10.1111/cogs.12015

Duyck, W., Vanderelst, D., Desmet, T., & Hartsuiker, R. J. (2008). The frequency effect in second-language visual word recognition. *Psychonomic Bulletin & Review*, *15*(4), 850–855. https://doi.org/10.3758/PBR.15.4.850

Einhäuser, W. (2017). The pupil as marker of cognitive processes. In Q. Zhao (Ed.), *Computational and cognitive neuroscience of vision* (pp. 141–169). Springer Science. https://doi.org/10.1007/978-981-10-0213-7_7

Francis, A. L., Tigchelaar, L. J., Zhang, R., & Zekveld, A. A. (2018). Effects of second language proficiency and linguistic uncertainty on recognition of speech in native and nonnative competing speech. *Journal of Speech, Language, and Hearing Research*, *61*(7), 1815–1830. https://doi.org/10.1044/2018_JSLHR-H-17-0254

Friedman, D., Hakerem, G., Sutton, S., & Fleiss, J. L. (1973). Effect of stimulus uncertainty on the pupillary dilation response and the vertex evoked potential. *Electroencephalography and Clinical Neurophysiology*, *34*(5), 475–484. https://doi.org/10.1016/0013-4694(73)90065-5

Garcia, D. L., & Gollan, T. H. (2022). The MINT sprint: Exploring a fast administration procedure with an expanded multilingual naming test. *Journal of the International Neuropsychological Society*, *28*(8), 845–861. https://doi.org/10.1017/S1355617721001004

Garcia Lecumberri, M. L., Cooke, M., & Cutler, A. (2010). Nonnative speech perception in adverse conditions: A review. *Speech Communication*, *52*(11–12), 864–886. https://doi.org/10.1016/j.specom.2010.08.014

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279. https://doi.org/10.1037/0033-295X.105.2.251

Goldinger, S. D., Luce, P. A., & Pisoni, D. B. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language*, *28*(5), 501–518. https://doi.org/10.1016/0749-596X(89)90009-0

Gollan, T. H., Montoya, R. I., Cera, C., & Sandoval, T. C. (2008). More use almost always means a smaller frequency effect: Aging, bilingualism, and the weaker links hypothesis. *Journal of Memory and Language*, *58*(3), 787–814. https://doi.org/10.1016/j.jml.2007.07.001

Haro, J., Guasch, M., Vallès, B., & Ferré, P. (2017). Is pupillary response a reliable index of word recognition? Evidence from a delayed lexical decision task. *Behavior Research Methods*, *49*(5), 1930–1938. https://doi.org/10.3758/s13428-016-0835-9

Herrmann, B., & Johnsrude, I. S. (2020). A model of listening engagement (MoLE). *Hearing Research*, *397*, 108016. https://doi.org/10.1016/j.heares.2020.108016

Hershman, R., Milshtein, D., & Henik, A. (2023). The contribution of temporal analysis of pupillometry measurements to cognitive research. *Psychological Research*, *87*(1), 28–42. https://doi.org/10.1007/s00426-022-01656-0

Hoeks, B., & Levelt, W. J. M. (1993). Pupillary dilation as a measure of attention: A quantitative system analysis. *Behavior Research Methods, Instruments, & Computers*, *25*(1), 16–26. https://doi.org/10.3758/BF03204445

Howes, D. (1957). On the relation between the intelligibility and frequency of occurrence of English words. *The Journal of the Acoustical Society of America*, *29*(2), 296–305. https://doi.org/10.1121/1.1908862

JASP Team. (2024). *JASP* (Version 0.19.0) [Computer software].

Kahneman, D. (1973). *Attention and effort*. Prentice-Hall.

Kapnoula, E. C., Jevtović, M., & Magnuson, J. S. (2024). Spoken word recognition: A focus on plasticity. *Annual Review of Linguistics*, *10*(1), 233–256. https://doi.org/10.1146/annurev-linguistics-031422-113507

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The sketch engine: Ten years on. *Lexicography*, *1*(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9

Kramer, S. E., Lorens, A., Coninx, F., Zekveld, A. A., Piotrowska, A., & Skarzynski, H. (2013). Processing load during listening: The influence of task characteristics on the pupil response. *Language and Cognitive Processes*, *28*(4), 426–442. https://doi.org/10.1080/01690965.2011.642267

Książek, P., Zekveld, A. A., Wendt, D., Fiedler, L., Lunner, T., & Kramer, S. E. (2021). Effect of speech-to-noise ratio and luminance on a range of current and potential pupil response measures to assess listening effort. *Trends in Hearing*, *25*. https://doi.org/10.1177/23312165211009351

Kuchinke, L., Võ, M. L.-H., Hofmann, M., & Jacobs, A. M. (2007). Pupillary responses during lexical decisions vary with word frequency but not emotional valence. *International Journal of Psychophysiology*, *65*(2), 132–140. https://doi.org/10.1016/j.ijpsycho.2007.04.004

Kuchinsky, S. E., Razeghi, N., & Pandža, N. B. (2023). Auditory, lexical, and multitasking demands interactively impact listening effort. *Journal of Speech, Language, and Hearing Research*, *66*(10), 4066–4082. https://doi.org/10.1044/2023_JSLHR-22-00548

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*(1), 1–36. https://doi.org/10.1097/00003446-199802000-00001

Marslen-Wilson, W. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*(1-2), 71–102. https://doi.org/10.1016/0010-0277(87)90005-9

Marslen-Wilson, W. (1991). Activation, competition, and frequency in lexical access. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 148–172). MIT Press.

Mathôt, S., & Vilotijević, A. (2023). Methods in cognitive pupillometry: Design, preprocessing, and statistical analysis. *Behavior Research Methods*, *55*(6), 3055–3077. https://doi.org/10.3758/s13428-022-01957-7

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, *27*(7–8), 953–978. https://doi.org/10.1080/01690965.2012.705006

McLaughlin, D. J., Zink, M. E., Gaunt, L., Spehar, B., Van Engen, K. J., Sommers, M. S., & Peelle, J. E. (2022). Pupillometry reveals cognitive demands of lexical competition during spoken word recognition in young and older adults. *Psychonomic Bulletin and Review*, *29*(1), 268–280. https://doi.org/10.3758/s13423-021-01991-0

Mor, B., & Prior, A. (2020). Individual differences in L2 frequency effects in different script bilinguals. *International Journal of Bilingualism*, *24*(4), 672–690. https://doi.org/10.1177/1367006919876356

Norris, D., McQueen, J. M., & Cutler, A. (2016). Prediction, Bayesian inference and feedback in speech recognition. *Language, Cognition and Neuroscience*, *31*(1), 4–18. https://doi.org/10.1080/23273798.2015.1081703

Peelle, J. E. (2018). Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior. *Ear and Hearing*, *39*(2), 204–214. https://doi.org/10.1097/AUD.0000000000000494

Pierrehumbert, J. B. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech*, *46*(2-3), 115–154. https://doi.org/10.1177/00238309030460020501

Pisoni, D. B. (2021). Cognitive audiology: An emerging landscape in speech perception. In J. Pardo, L. Nygaard, R. Remez, & D. B. Pisoni (Eds.), *The handbooks of speech perception* (pp. 697–730). Wiley-Blackwell.

Poeppel, D., & Sun, Y. (2021). Neural encoding of speech and word forms. In A. Papafragou, J. C. Trueswell, & L. R. Gleitman (Eds.), *The Oxford handbook of the mental lexicon* (pp. 53–73). Oxford University Press.

Pollack, I., Rubenstein, H., & Decker, L. (1959). Intelligibility of known and unknown message sets. *The Journal of the Acoustical Society of America*, *31*(3), 273–279. https://doi.org/10.1121/1.1907712

R Core Team. (2024). *A language and environment for statistical computing* (4.3.3). R Foundation for Statistical Computing. https://www.r-project.org/

Richer, F., & Beatty, J. (1987). Contrasting effects of response uncertainty on the task-evoked pupillary response and reaction time. *Psychophysiology*, *24*(3), 258–262. https://doi.org/10.1111/j.1469-8986.1987.tb00291.x

Rojas, C., Vega-rodríguez, Y. E., Lagos, G., & Crisosto-alarcón, J. (2024). Applicability and usefulness of pupillometry in the study of lexical access. A scoping review of primary research. *Frontiers in Psychology*, *15*(1372912). https://doi.org/10.3389/fpsyg.2024.1372912

Rönnberg, J., Signoret, C., Andin, J., & Holmer, E. (2022). The cognitive hearing science perspective on perceiving, understanding, and remembering language: The ELU model. *Frontiers in Psychology*, *13*(September), 1–17. https://doi.org/10.3389/fpsyg.2022.967260

Savin, H. (1963). Word-frequency effect and errors in the perception of speech. *The Journal of the Acoustical Society of America*, *35*(2), 200–206. https://doi.org/10.1121/1.1918432

Scharenborg, O., & van Os, M. (2019). Why listening in background noise is harder in a non-native language than in a native language: A review. *Speech Communication*, *108* (August 2018), 53–64. https://doi.org/10.1016/j.specom.2019.03.001

Schmidtke, J. (2014). Second language experience modulates word retrieval effort in bilinguals: Evidence from pupillometry. *Frontiers in Psychology*, *5*(137). https://doi.org/10.3389/fpsyg.2014.00137

Schmidtke, J. (2016). The bilingual disadvantage in speech understanding in noise is likely a frequency effect related to reduced language exposure. *Frontiers in Psychology*, *7*(678), 1–15. https://doi.org/10.3389/fpsyg.2016.00678

Schmidtke, J. (2018). Pupillometry in linguistic research: An introduction and review for second language researchers. *Studies in Second Language Acquisition*, *40*(3), 529–549. https://doi.org/10.1017/S0272263117000195

Schmidtke, J., & Tobin, S. (2024). Pupil dilation as a dependent variable in language research. In M. H. Papesh & S. D. Goldinger (Eds.), *Modern pupillometry* (pp. 201–227). Springer. https://doi.org/10.1007/978-3-031-54896-3_7

Seabold, S., & Perktold, J. (2010, June 28–July 3). *Statsmodels: Econometric and statistical modeling with python*. 9th python in science conference.

Shi, L.-F. (2014). Lexical effects on recognition of the NU-6 words by monolingual and bilingual listeners. *International Journal of Audiology*, *53*(5), 318–325. https://doi.org/10.3109/14992027.2013.876109

Shi, L.-F. (2015). English word frequency and recognition in bilinguals: Inter-corpus comparison and error analysis. *International Journal of Audiology*, *54*(10), 674–681. https://doi.org/10.3109/14992027.2015.1030509

Strauch, C., Wang, C. A., Einhäuser, W., Van der Stigchel, S., & Naber, M. (2022). Pupillometry as an integrated readout of distinct attentional networks. *Trends in Neurosciences*, *45* (8), 635–647. https://doi.org/10.1016/j.tins.2022.05.003

Strauß, A., Wu, T., McQueen, J. M., Scharenborg, O., & Hintz, F. (2022). The differential roles of lexical and sublexical processing during spoken-word recognition in clear and in noise. *Cortex*, *151*, 70–88. https://doi.org/10.1016/j.cortex.2022.02.011

Tanenhaus, M. K., Magnuson, J. S., Dahan, D., & Chambers, C. G. (2000). Eye movements and lexical access in spoken-language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing. *Journal of Psycholinguistic Research*, *29*(6), 557–580. https://doi.org/10.1023/A:1026464108329

Van Engen, K. J., Dey, A., Runge, N., Spehar, B., Sommers, M. S., & Peelle, J. E. (2020). Effects of age, word frequency, and noise on the time course of spoken word recognition. *Collabra: Psychology*, *6*(1), 1–10. https://doi.org/10.1525/collabra.17247

van Rij, J., Hendriks, P., van Rijn, H., Baayen, R. H., & Wood, S. N. (2019). Analyzing the time course of pupillometric data. *Trends in Hearing*, *23*, 1–22. https://doi.org/10.1177/2331216519832483

Whitford, V., & Titone, D. (2012). Second-language experience modulates first- and second-language word frequency effects: Evidence from eye movement measures of natural paragraph reading. *Psychonomic Bulletin & Review*, *19*(1), 73–80. https://doi.org/10.3758/s13423-011-0179-5

Winsler, K., Midgley, K. J., Grainger, J., & Holcomb, P. J. (2018). An electrophysiological megastudy of spoken word recognition. *Language, Cognition and Neuroscience*, *33*(8), 1063–1082. https://doi.org/10.1080/23273798.2018.1455985

Zekveld, A. A., Koelewijn, T., & Kramer, S. E. (2018). The pupil dilation response to auditory stimuli: Current state of knowledge. *Trends in Hearing*, *22*, 233121651877717. https://doi.org/10.1177/2331216518777174

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2010). Pupil response as an indication of effortful listening: The influence of sentence intelligibility. *Ear and Hearing*, *31*(4), 480–490. https://doi.org/10.1097/AUD.0b013e3181d4f251