The Version of Record of this manuscript has been published and is available in INTERNATIONAL JOURNAL OF AUDIOLOGY, 28/07/2020

https://www.tandfonline.com/doi/full/10.1080/14992027.2020.1798518

Design and evaluation of the effectiveness of a corpus of congruent and incongruent English sentences for the study of event related potentials

Joaquin T. Valderrama [a,b,*] Elizabeth F. Beach [a], Mridula Sharma [b], Shivali Appaiah-Konganda [b], Elaine Schmidt [b,c].

[a] *National Acoustic Laboratories, Australian Hearing Hub. 16 University Avenue, Macquarie University, New South Wales, 2109, Sydney, Australia.*

[b] *Department of Linguistics, Australian Hearing Hub. 16 University Avenue, Macquarie University, New South Wales, 2109, Sydney, Australia.*

[c] *Department of Theoretical and Applied Linguistics, University of Cambridge. 9 West Road, Cambridge, CB3 9DP, United Kingdom.*

* Corresponding author

Joaquin T. Valderrama
National Acoustic Laboratories
Australian Hearing Hub
Level 5, 16 University Avenue
Macquarie University NSW 2109
Sydney, Australia
Phone: +61 2 9412 6878
Email address: joaquin.valderrama@nal.gov.au, joaquin.valderrama@mq.edu.au.
ORCID: 0000-0002-5529-8620

Abstract

Objective: To design and evaluate the effectiveness of a stimulus material in eliciting the N400 event related potential (ERP).

Design: A set of 700 semantically congruent and incongruent sentences was developed in accordance with current linguistic norms, and validated with an electroencephalography (EEG) study, in which the influence of age and gender on the N400 ERP magnitude was analysed.

Study sample: Forty-five normal-hearing subjects (19-57 years, 21 females) participated in the EEG study.

Results: The stimulus material used in the EEG study elicited a robust N400 ERP, with a morphology consistent with the literature. Results also showed no statistically significant effect of age or gender on the N400 magnitude.

Conclusions: The material presented in this paper constitutes the largest complete stimulus set suitable for both auditory and text-based N400 experiments. This material may help facilitate the efficient implementation of future N400 ERP studies, as well as promote standardization and consistency across studies.

Keywords

N400; Speech Perception; Language-Related ERPs; Semantic Violation.

1. Introduction

The analysis of event-related potentials (ERPs) has been fundamental for understanding the neural basis of language encoding in the human brain. In particular, the N400 ERP has been identified as an index of lexical and semantic processing (Kutas and Hillyard, 1980; Osterhout and Holcomb, 1992; Brown and Hagoort, 1993, Hagoort, 2008), and is widely used in the design of experiments aiming to evaluate language comprehension (Boudewyn et al., 2012; Federmeier, 2007; Kutas and Federmeier, 2000; Van Petten and Luka, 2012). The N400 is characterized as a negative deflection of the average ERP waveform that peaks around 400 ms after a semantic violation, presenting a larger magnitude at midline central-parietal sites (Kutas, 1993; Duncan et al., 2009; Jerger and Martin, 2009).

The large body of literature around the N400 has confirmed that this ERP is associated with language processing (Kutas and Federmeier, 2000; Federmeier, 2007). In particular, a review of studies using electrophysiology, functional magnetic resonance imaging, magnetoencephalography, and intracranial recordings showed that the middle temporal gyrus is the main area of the brain involved in generation of the N400 (Lau et al., 2008), an area involved in long-term storage of lexico-semantic information (Hickok and Poeppel, 2004, 2007; Gitelman et al., 2005; Martin, 2007).

A typical N400 ERP study involves the auditory and/or visual presentation of a number of congruent and incongruent sentences[1], and the recording of the subject´s associated neural response through surface electrodes placed on the head. The N400 is estimated by comparing the average ERP waveforms corresponding to the congruent and incongruent sentences. It is well known that the creation of the stimulus material is a complex and arduous process that includes: (i) ideation of a set of congruent and incongruent sentences with a consistent morpho-

[1] The semantic violation that elicits the N400 ERP can also be primed by a list of words that do not form sentences or discourse (Titone and Salisbury, 2004; Romei et al. 2011; Brown and Hagoort, 1993).

syntactic structure; (ii) evaluation of the sentences by independent reviewers; and in case of auditory stimuli, (iii) recording of the sentences by a trained speaker using sophisticated equipment such as a studio microphone, a high-fidelity soundcard, and a sound-proof booth; and (iv) advanced post-processing of the audio files (Duncan et al., 2009; Swaab et al., 2011).

Most N400 studies use stimulus materials that have been developed in-house, and often the sentences are provided as an appendix (e.g. Litcofsky and Van Hell, 2017; Holt et al., 2018). However, sentence development is very time-consuming and with multiple researchers repeating the process, it is not only highly inefficient, it makes comparison of results difficult. Varying degrees of semantic violation in sentences can lead to variation the amplitude of the elicited N400 response (Kutas and Hillyard, 1984), which further adds to the difficulty of comparing results between different studies that use different stimuli. If researchers had access to a standardized set of sentences, this would reduce the need for repeated stimulus development by different research groups and also facilitate comparisons between different studies (Bradshaw, 1984). In 1980, Bloom and Fischler published one of the most widely used sentence sets. This study provided 329 sentences with different levels of predictability. A subset of sentences from this study was used by Kutas and Hillyard (1984) to confirm that the magnitude of the N400 was inversely correlated with the degree of expectation associated with the final word of the sentences. Although Bloom and Fischler (1980) provided 329 sentences, studies that evaluate multiple different test conditions often require a larger number of stimuli. Block and Baldwin (2010) extended Bloom and Fischler's set by adding 398 new sentences that followed a similar format. Although this expanded sentence set is likely to be large enough for visual-only studies in which sentences are presented as text only, each N400 study that presented auditory or auditory-visual stimuli would still need to make a recording of their selected sentences for use in their study. To the best of our knowledge, no researchers have published all of the necessary stimulus material to run an N400 ERP study with auditory stimuli.

This paper provides a large set of congruent and incongruent English sentences, details the creation process, evaluates the quality of the sentences through a subjective evaluation of their meaningfulness, and provides the stimulus material – in text format and auditory recordings. In addition, the paper presents the results of an electrophysiology validation study which assessed the appropriateness of the stimulus material for N400 ERP studies. This involved evaluation of the N400 through grand-average ERP signals, analysis of the scalp-distribution, characterization of the individual variability of its magnitude, and an investigation of the influence of age and gender in a large set of normal hearing subjects.

2. Stimulus material

2.1. Congruent and incongruent sentences

A set of 350 congruent and 350 incongruent sentences was created according to the following structure: <<The + [1st noun: 2 syllables] + [verb: 1 syllable] + the + [2nd noun: 2 syllables] + [complement: 3 syllables]>>. An example of a sentence with this structure could be 'The toddler likes the biscuits with some milk'. The first noun and the verb defined the context of the sentence, whereas the congruency of the sentence was determined by the second noun, i.e. the 'critical word'. Incongruent sentences were those in which the critical word was not coherent with the context. The final complement provided the sentence with continuity following the potential congruency violation in order to avoid an overlap of N400 and wrap-up effects (Hagoort and Brown, 2000). The preposition <<the>> before the critical word aimed to provide an identical preceding phonetic context for all critical words in order to standardize the assimilation effect (i.e. a change in the pronunciation of a phoneme due to an adjacent sound) across the entire set of sentences (Ohala, 1988).

[Table 1]

Congruent sentences can be recorded naturally, however, reading aloud incongruent sentences can lead to acoustic confounds derived from exaggerated or unnatural prosody introduced by

the speaker who is aware that what they are saying is somehow 'odd' (Dimitrova et al., 2012; Meulman et al., 2014). Thus, it is conceivable that listeners might have access to prosodic cues indicating that a violation is to come before it has actually occurred. This expectancy would skew any N400 effects. To overcome this confound, each incongruent sentence was constructed by combining two naturally spoken congruent sentences using a cross-splicing procedure (Steinhauer et al., 2010; Meulman et al., 2014). For each incongruent sentence, an additional sentence was created by replacing the incongruent critical word with a different noun that was congruent with the context of the sentence. Table 1 shows an example of the three types of sentences created. In this example, sentences 1a and 1c are congruent sentences that were recorded naturally. Sentence 1b is an incongruent sentence that was constructed by cross-splicing the critical word from sentence 1c ('glasses') and replacing it with 'petrol' from 1a. To facilitate the cross-splicing procedure, we chose critical words with phonemes that have clear onsets following a brief silence in the otherwise continuous speech stream (stops: [p], [b], [t], [k], [g]; affricates: [tʃ] or [dʒ]); and avoided words that started with vowels.

All words used in the sentences were contained in SUBTLEX-UK, a word-frequency database for British English based on subtitles of British television programmes (Van Heuven et al., 2014). In order to ensure that the words of the sentences were familiar to the subjects, only words with a *Zipf* value[2] between 4 and 7 were selected. This value is an indication of medium- to high- word frequency (Monsell et al., 1989; Van Heuven et al, 2014).

[Figure 1, single column]

The meaningfulness of the congruent and incongruent sentences was evaluated by young adult native English-speaking students from Macquarie University (Sydney, Australia). The

[2] The *Zipf scale* takes the log10 of the frequency per billion words (Van Heuven et al., 2014). Thus a *Zipf* value equal to 4 indicates a frequency per million words (*fpmw*) of 10; and a *Zipf* equal to 7 corresponds to a *fpmw* of 10,000. This scale comes from the American linguist George Kingsley Zipf, who was the first to formulate a law about the regularities of word frequency distribution (Zipf, 1949).

participants read each sentence and provided a rating between 1 and 6, where 1 – 'The sentence makes complete sense' and 6 – 'The sentence makes no sense at all'. Each sentence was evaluated by at least five evaluators. The presentation order of the 700 sentences was randomized, and a final score was obtained for each sentence by estimating the mean of the evaluations. Sentences with a mean score closer to 1 were considered very congruent, while sentences with a mean score closer to 6 were considered very incongruent. During a final review of the list of sentences, five incongruent sentences were found to be potentially congruent, and therefore, they were assigned a congruency score of 3.5. Figure 1 shows the histogram of the subjective meaningfulness ratings of the congruent and incongruent sentences. This figure shows that the distribution is scattered towards the extreme values, indicating a predominance of sentences rated by the evaluators as either very congruent or very incongruent.

2.2. Questions and fillers

A set of questions that focused on the content of the sentences was also created. The questions occurred randomly during stimulus presentation to help sustain the attention of the participants. This set consisted of 200 *wh*-questions with an equal number focused on the congruent and incongruent sentences. For the sentences shown in table 1, the associated questions (Q) and responses (R) were [congruent sentence] Q: *Where does the driver put the petrol?* R: *In the car*; [incongruent sentence] Q: *Where does the mother break the petrol?* R: *On the shelf*. In addition, 130 filler sentences were devised to reduce predictability. These were similar in duration to the stimulus sentences, but differed in structure. Some examples of these fillers were "*Every Sunday the father goes to church*" and "*The package was sent by express post*".

2.3. Audio recordings

The sentences, questions and fillers were recorded by a trained female native Australian English speaker in a sound-proof recording studio using a C535-EB vocal microphone (AKG Acoustics GmbH, Vienna, Austria), a StudioLive 16.4.2 audio mixer (PreSonus, Baton Rouge, LA), and a

sampling rate of 48 kHz. The congruent sentences (i.e. XXXa) and the additional congruent sentences necessary to form the incongruent sentences (i.e. XXXc) were recorded consecutively at least twice. To facilitate cross-splicing, the speaker was instructed to emphasize the critical word and to give a brief pause of about 0.5 seconds after the end of the critical word.

The recorded audio files were processed offline using Praat (Boersma, 2001; Boersma and Weenink, 2016). Processing consisted of (1) extracting the sentences from the continuous audio files; (2) selecting the best candidate from the recorded options for each sentence based on intonation, creakiness, clarity, intensity, and ease for cross-splicing; (3) cross-splicing the critical word to construct the incongruent sentences (i.e., XXXb); (4) adjusting the intensity of all sentences according to their root-mean square (RMS) value; and (5) setting the time points at which relevant language components occurred (markers) in order to identify the onset of the associated language-related ERPs during subsequent data processing. Markers were placed at (i) the onset of the initial <<The>>, (ii) the onset of the <<the>> preceding the critical word, (iii) the onset of the critical word, (iv) the onset of the complement, and (v) the end of the sentence. All cross-splicing edits were performed at zero-crossing points to avoid undesired audible artefacts like clicks or pops. The quality of the final audio files and the absence of imperfections in the cross-splicing process was validated by author 5.

The full list of congruent and incongruent sentences, questions, and fillers, along with the mean subjective evaluation of their meaningfulness, are provided as supporting material in appendix A. This appendix also includes a description of the rationale for the excluded sentences, i.e. those with a congruency rating set to 3.5. The raw audio files and markers are also provided as supplementary material (appendix B).

3. Experimental validation

The feasibility of the proposed set of sentences to evoke the N400 ERP was evaluated with an electroencephalography (EEG) study, in which the influence of age and gender on the N400 magnitude was analysed.

3.1. Methods

3.1.1. Ethics

The experimental protocol followed in this study was in accordance with the National Statements on Ethical Conduct in Human Research and was approved by the Human Research Ethics Committees of Macquarie University and Australian Hearing (Refs 5201400862; AHHREC2014-5).

3.1.2. Participants

Forty-five participants (aged 19-57, mean = 38.78 years, SD = 11.22 years, 21 females) were recruited from the general community and Macquarie University. The inclusion criteria required that participants had English as a first language and normal or near-normal pure-tone hearing thresholds in both ears in the typical range of frequencies evaluated in the clinic (Dillon, 2012; Katz, 2014). Normal hearing was defined as a hearing loss ≤ 20 dB hearing level (HL) at 0.25 – 6 kHz; and near-normal thresholds were considered as ≤ 25 dB HL up to 2 kHz, ≤ 30 dB HL at 3 kHz, ≤ 35 dB HL at 4 kHz, and ≤ 40 dB HL at 6 kHz (Moore et al., 2012). All participants gave written consent to participate, and received $40 after completing the study.

3.1.3. Auditory stimulus

The auditory stimuli consisted of 80 congruent sentences, 80 incongruent sentences, 14 fillers, and 24 questions taken from the recorded materials described in section 2. Eighty highly congruent and 80 highly incongruent sentences were chosen randomly from the 160 sentences with the lowest and highest meaningfulness score, i.e. the 160 most- and 160 least-congruent sentences respectively. All of the 160 most congruent sentences were rated as 1.0, and the ratings for the 160 least congruent sentences were between 5.8 and 6.0. The 14 fillers were

selected randomly from the list of fillers and presented randomly every 6 to 10 sentences, i.e. with a probability of occurrence of 8.7%. The 24 questions were equally distributed between the congruent and incongruent sentences, and were presented randomly every 2 to 11 sentences, i.e. with a probability of occurrence of 15%. A short 1 kHz tone (or 'beep') was presented 1 second before every question to inform the participants that they were about to hear a question about the preceding sentence and that a brief oral response was expected from them. The time between the end of one sentence and the onset of the following sentence was 4 seconds, except for questions, in which a time period of 8 seconds was provided to allow participants time to respond. The sentences, fillers and questions were distributed in four blocks of about 7 minutes each.

3.1.4. Stimulus presentation

The auditory stimuli were presented to the subjects diotically at 60 dB sound-pressure level (SPL) through ER-3A insert earphones (Etymotic Research Inc., Elk Grove Village, IL) placed in the ear canal after otoscopic examination. The insert earphones were connected to a Fireface UCX audio soundcard (RME Audio, Haimhausen, Germany). Stimulus level was calibrated in a type HA2 artificial ear 2-cc acoustic coupler, connected to a type 4144 pressure microphone, which was connected to a type 2636 measuring amplifier through a type 2639 preamplifier cable (Brüel & Kjæer Sound & Vibration Measurement A/S, Nærum, Denmark).

3.1.5. EEG recording

The EEG recording sessions took place in an electromagnetically shielded booth at the Speech and Hearing Clinic of Macquarie University. During the session, participants were seated in a comfortable armchair and were instructed to remain still during the test, to leave their neck and shoulder muscles relaxed, and to respond orally to the questions that followed a beep sound.

The recording of the neural response was carried out by 64 surface electrodes placed on the head using the TC64 EasyCap EEG recording cap (EASYCAP GmbH, Herrsching, Germany), in

which the electrode positions were distributed according to the International 10-20 system convention (Klem et al., 1999). The ground electrode was placed at AFz, and all active electrodes were referenced to the left mastoid (i.e., Tp9). The impedances of the electrodes with the skin were kept below 5 kΩ in all recordings. The EEGs were acquired by a Neuroscan SynAmps RT recording system, which was controlled by Curry 7 software (Compumedics Limited, Abbotsford, Australia). The recording sampling rate was 1 kHz; and the bandpass of the analogue filters was [0.01 - 300] Hz.

3.1.6. Data analysis

EEG data were analysed with custom scripts developed in Matlab (The Mathworks Inc., Natick, MA). The EEG recordings were processed in a 4-step procedure: (1) re-referencing, (2) digital filtering, (3) blink-artifact suppression, and (4) segments averaging. First, EEGs were re-referenced to the combined mastoid by subtracting from each raw EEG channel (i.e., [XX-Tp9]) the Tp10 EEG channel divided by 2, i.e. [Tp10-Tp9]/2. This way, all EEG channels were referenced to the combined mastoid, as [XX-Tp9] – [Tp10-Tp9]/2 = [XX – (Tp9+Tp10)/2]. The re-referenced EEGs were digitally filtered by a zero-phase 4^{th} order Butterworth filter with bandpass cut-off frequencies [0.05 - 20] Hz. Blink artifacts were suppressed using iterative template matching and suppression (ITMS: Valderrama et al., 2018), a technique that allows blink-artifact suppression from single-channel EEG recordings. In the present multichannel application, ITMS was first applied to FP1 (i.e., an EEG channel situated in the vertical of the left eye) in order to facilitate detection of blink events. Once blink events were detected on the FP1 channel, a simplified version of ITMS was applied to the remaining EEG channels. The simplified ITMS procedure consisted of the following processes [described in detail in (Valderrama et al., 2018)]: (1) template estimation, (2) amplitudes estimation, (3) blink-artifact model estimation, and (4) blink-artifact model suppression. The Matlab functions that implement the original and simplified versions of ITMS are provided as supporting material (appendix C). Finally, the EEG

segments corresponding to 1 second pre- and 3 seconds post- critical word onset were averaged to obtain the ERPs associated with congruent and incongruent sentences.

The N400 ERP was quantified according to the area under the curve (AuC), which was estimated as the area (in s·μV) between the congruent and incongruent ERPs in the [0.4 – 0.8] s time interval (Swaab et al., 2011). Grand-average ERP signals were obtained by averaging the ERPs associated with congruent and incongruent sentences across participants. For this analysis, the 10% of the ERPs with highest root-mean square (RMS) value in each channel (i.e., those most contaminated by noise) were discarded from the average in order to improve the quality of the ERPs (Thornton, 2007). Scalp topographic maps were also created using functions from the FieldTrip Matlab toolbox (Oostenveld et al., 2011). In addition, a cluster analysis was carried out to evaluate the N400 ERP in the frontal, central, and parietal-occipital areas of the brain. In each area, clusters were formed by averaging the ERPs corresponding to the following channels: frontal [AF3, AF1, AF2, AF4, F3, F1, Fz, F2, F4, FC3, FC1, FCz, FC2, FC4]; central [FC3, FC1, FCz, FC2, FC4, C3, C1, Cz, C2, C4, CP3, CP1, CPz, CP2, CP4]; and parietal-occipital [CP3, CP1, CPz, CP2, CP4, P3, P1, Pz, P2, P4, PO3, PO4, O1, Oz, O2].

The cluster analysis evaluated the presence of the N400 ERP, and assessed the influence of gender and age on its magnitude. The presence of the N400 ERP was evaluated by a one-sample *t*-test, in which the null hypothesis was that the mean of the AuC distribution was zero (i.e., no N400 ERP was obtained). The gender analysis was based on a two-sample *t*-test, in which the null hypothesis was that the mean of the AuC distribution for males and females were equal (i.e., males and females showed similar AuCs). The influence of age on the AuC was evaluated using a linear regression test, in which the null hypothesis was that the slope of the correlation was equal to zero (i.e., AuC was not influenced by the age of the participants).

Statistical analysis was conducted using functions from the 'Statistics and Machine Learning' Matlab toolbox. The normality assumption was evaluated in all distributions according to the

Lilliefors test of normality; and a conservative alpha-value of 0.01 was chosen as the criterion for statistical significance.

3.2. Results

3.2.1. Hearing thresholds

[Figure 2, double column]

Figure 2 shows the pure-tone audiometric threshold distributions for the left and right ears at the test frequencies. Boxplots represent the quartiles of the distribution. This figure shows that of the 45 participants, 44 presented with normal or near-normal hearing according to the study criteria. The remaining participant met the near-normal criteria for all but one of their thresholds, which was slightly outside the inclusion criteria.

3.2.2. N400 ERP

[Figure 3, double column]

Figure 3.A shows the grand-average ERPs associated with the onset of the congruent (in blue) and incongruent (in brown) critical words at selected EEG channels. The N400 ERP is shown in this figure as a negative voltage deviation of the incongruent sentence compared to the congruent sentence in the time interval [0.4 – 0.8] s after the critical word onset. The topographic maps presented in figure 3.B show that the N400 ERP magnitude is highest in the central area of the brain around the aforementioned time interval.

[Figure 4, double column]

Figure 4 presents an analysis of the N400 AuC at the frontal, central, and parietal-occipital areas. The first column shows the grand-average ERPs corresponding to each cluster of EEG channels. The second column shows the AuC histogram, the fitted distribution, and the probability (p) that the AuC distribution comes from a normal distribution with mean equal to 0. The results presented in the three evaluated areas show that all the p-values were statistically significant,

thus indicating that the set of sentences used in this study successfully evoked the N400 ERP. However, it is also noteworthy that a significant portion of the participants presented negative AuC values in each of the three areas [frontal: 14 (31%), central: 10 (22%), parietal-occipital: 9 (20%)]. The fitted distributions for males and females presented in the third column show that females tended to show greater AuC, although this difference was not significant in any of the three topographic areas. Finally, the age plots shown in the fourth column indicate no effect of age on the AuC estimate.

4. Discussion

This paper is an important contribution to the N400 literature, providing researchers with access to a new verified set of 700 congruent and incongruent English sentences and their audio recordings, appropriate for use in N400 ERP studies. To our knowledge, this constitutes the largest complete stimulus set suitable for both auditory and text-based experiments.

The subjective evaluation of the meaningfulness of the sentences showed that most sentences were classified as either highly congruent (on a 1-to-6 scale, 304/350 sentences had a mean score less than or equal to 1.5) or highly incongruent (250/350 sentences had a mean score greater than or equal to 5.5). This large set of highly congruent and incongruent sentences may be useful when testing subjects in multiple sessions or in different stimulus conditions [e.g. different SNRs, transducers (insert vs loudspeakers), etc.].

The ability of the sentences to elicit the N400 ERP was evaluated in a sample of 45 subjects with good hearing. In order to test a significant portion of the proposed set, this experiment used 80 sentences selected randomly from the subset of the 160 most congruent sentences, and 80 sentences from the 160 most incongruent sentences. This experiment demonstrated the suitability of the selected test sentences for evoking N400 ERP. Consistent with the majority of published N400 studies (Osterhout and Nicol, 1999; Swaab et al., 2011; Boudewyn et al., 2012), the grand-average ERPs and the topographic maps presented in figure 3 showed a dominant

central distribution of the N400 magnitude. In addition, the incongruous sentences produced an ERP that was more positive at the peak that followed the N400 in the parietal area. This effect is known as 'post-N400 positivity' (Matsumoto et al., 2005; Van Petten and Luka, 2006; Federmeier et al., 2007), and is associated with an impossible or semantically anomalous interpretation (Van Petten and Luka, 2012).

The distributions of the individual AuC estimates at the midline frontal, central and parietal areas had statistically significant positive means, thus providing evidence of the elicitation of the N400 ERP at a group level. However, these distributions also showed that a significant portion of the individuals [frontal: 14 subjects (31%); central: 10 subjects (22%); parietal: 9 subjects (20%)] presented negative AuC values, consistent with other studies showing only limited reliability of N400 ERP studies when conducting single-subject analyses (Cruse et al., 2014).

The analysis also showed that although females presented fitted AuC distributions with a higher mean, this effect was not statistically significant. The few previous studies that have analysed the influence of gender on the N400 ERP have reported conflicting results. On the one hand, Tsolaki et al. (2015) showed no gender-related differences; whereas several other studies have found that females presented earlier and larger N400 ERP components (Daltrozzo et al., 2007; Proverbio et al., 2010; Steffensen et al., 2008; Wirth et al., 2007). Whether or not males and females process lexical violations differently is still under debate. With regard to age, we observed no effect on the AuC in any of the midline clusters. These results are consistent with recent literature, in which no significant N400 changes with age have been reported (Federmeier et al., 2003; Grieder et al., 2012; Komes et al., 2014; Tsolaki et al., 2015; Wilkinson et al., 2013). It is noteworthy that the age and gender analyses conducted in this paper were supported by a gender-balanced distribution of the participants (21 females against 24 males), and by a uniform distribution of age across a wide range (from 19 to 57 years).

Attached to this paper, the full text of all stimulus sentences, filler sentences, and questions are provided as supporting material (appendix A). In addition, this paper also provides the processed

audio files (appendix B). Access to this stimulus material is intended to facilitate the efficient implementation of future N400 ERP studies, and promote standardization and comparisons across studies.

Two important limitations should be considered when using the stimulus material presented in this paper for other studies. The first limitation is that the audio files were recorded by an Australian-English speaker, which could affect the results if a study is conducted in a population that uses a different English dialect. In this case, it is advised that the sentences are recorded by a trained speaker of the same dialect as the study population, following the steps described in section 2.3. The second limitation is that the subjective rating of the congruency of the sentences was carried out by young adult university students only, which may introduce bias since their evaluations are not being representative of the general population (which included individuals with a lower education level and from different age groups). In cases where the study population differs significantly from the university-educated young adults who rated the congruency of the sentences in this study, the authors suggest that the congruency of the sentences is re-evaluated by people with a similar profile as the study population.

Acknowledgements

The authors gratefully acknowledge Ms Lorna Betts for her help with the audio recording of the sentences; Mr Greg Stewart (NAL: National Acoustic Laboratories, Sydney, Australia) for his help with the calibration of the stimuli; and the Macquarie University students who participated in the subjective evaluation of the sentences. This work was supported by the Australian Government Department of Health. No potential conflict of interest was reported by the authors.

Appendix

Supplementary material associated with this article can be found at [URL]. Appendix A presents a table of (1) the congruent and incongruent sentences; (2) the associated individual and mean

subjective congruency ratings; (3) the list of questions and expected responses; (4) the list of filler sentences; and (5) the rationale for excluding the ambiguous sentences. Appendix B includes the processed audio files of the recorded sentences and markers. Appendix C includes the Matlab functions and associated files necessary to implement the blink-artifact removal technique 'iterative template matching and suppression (ITMS)' in a multichannel EEG configuration.

References

Block, C.K., Baldwin, C.L. (2010). Cloze probability and completion norms for 498 sentences: Behavioral and neural validation using event-related potentials. Behavior Research Methods 42, 665-670.

Bloom, P.A., Fischler, I. (1980). Completion norms for 329 sentence contexts. Memory & Cognition 8, 631-642.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. Glot International 5, 341-345.

Boersma, P., Weenink, D. (2016). Praat: doing phonetics by computer [Computer program]. Version 6.0.19, retrieved 13 June 2016 from http://www.praat.org/

Bradshaw, J.L. (1984). A guide to norms, ratings, and lists. Memory & Cognition 12, 202-206.

Brown, C., Hagoort, P.J. (1993). The processing nature of the N400: evidence from masked priming. Journal of Cognitive Neuroscience 5, 34-44.

Boudewyn, M.A., Gordon, P.C., Long, D., Polse, L., Swaab, T.Y. (2012). Does discourse congruence influence spoken language comprehension before lexical association? Evidence from event-related potentials. Language and Cognitive Processes 27, 698-733.

Cruse, D., Beukema, S., Chennu, S., Malins, J.G., Owen, A.M., McRae, K. (2014). The reliability of the N400 in single subjects: Implications for patients with disorders of consciousness. NeuroImage: Clinical 4, 788-799.

Daltrozzo, J., Wioland, N., Kotchoubey, B. (2007). Sex differences in two event-related potential components related to semantic priming. Archives of Sexual Behavior 36, 555-568.

Dillon, H. (2012). Hearing Aids (Thieme Publishers, New York), 608 p.

Dimitrova, D.V., Stowe, L.A., Redeker, G., Hoeks, J.C. (2012). Less is not more: neural responses to missing and superfluous accents in context. Journal of Cognitive Neuroscience 24, 2400-2418.

Duncan, C.C., Barry, R.J., Connolly, J.F., Fischer, C., Michie, P.T., Näätänen, R., Polich, J., Reinvang, I., Van Petten, C. (2009). Event-related potentials in clinical research: Guidelines for eliciting, recording, and quantifying mismatch negativity, P300, and N400. Clinical Neurophysiology 120, 1883-1908.

Federmeier, K.D. (2007). Thinking ahead: the role and roots of prediction in language comprehension. Psychophysiology 44, 491-505.

Federmeier, K.D., Van Petten, C., Schwartz, T.J., Kutas, M. (2003). Sounds, words, sentences: age-related changes across levels of language processing. Psychology and Aging 18, 858-872.

Federmeier, K.D., Wlotko, E.W., de Ochoa-Dewald, E., Kutas, M. (2007). Multiple effects of sentential constraint on word processing. Brain Research 1146, 75-84.

Gitelman, D.R., Nobre, A.C., Sonty, S., Parrish, T.B., Mesulam, M.M. (2005). Language network specializations: an analysis with parallel task designs and functional magnetic resonance imaging. Neuroimage 26, 975-985.

Grieder, M., Crinelli, R.M., Koenig, T., Wahlund, L.-O., Dierks, T., Wirth, M. (2012). Electrophysiological and behavioural correlates of stable automatic semantic retrieval in aging. Neuropsychologia 50, 160-171.

Hagoort, P. (2008). The fractionation of spoken language understanding by measuring electrical and magnetic brain signals. Philosophical Transactions of the Royal Society of London – Series B: Biological Sciences 363, 1055-1069.

Hagoort, P., Brown, C. M. (2000). ERP effects of listening to speech: semantic ERP effects. Neuropsychologia 38, 1518-1530.

Hickok, G., Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition 92, 67-99.

Hickok, G., Poeppel, D. (2007). The cortical organization of speech processing. Nature Reviews Neuroscience 8, 393-402.

Holt, R., Kung, C., Demuth, K. (2018). Listener characteristics modulate the semantic processing of native vs. foreign-accented speech. PLoS ONE 13(12): e0207452.

Jerger, J., Martin, J. (2009). Some effects of aging on event-related potentials during a linguistic monitoring task. International Journal of Audiology 44, 321-330.

Klem, G.H., Lüders, H.O., Jasper, H.H., Elger, C. (1999). The ten-twenty electrode system of the International Federation. The International Federation of Clinical Neurophysiology. Electroencephalography and Clinical Neurophysiology 52, 3-6.

Katz, J. (2014). Handbook of clinical audiology, edited by Marshall Chasin, Kristina English, Linda J. Hood, and Kim L. Tillery (Wolters Kluwer Health, Philadelphia), 927 p.

Komes, J., Schweinberger, S.R., Wiese, H. (2014). Fluency affects source memory for familiar names in younger and older adults: evidence from event-related brain potentials. NeuroImage 92C, 90-105.

Kutas, M. (1993). In the company of other words: electrophysiological evidence for single-word and sentence context effects. Language and Cognitive Process 8, 533-572.

Kutas, M., Federmeier, K.D. (2000). Electrophysiology reveals semantic memory use in language comprehension. Trends in Cognitive Sciences 4, 463-470.

Kutas, M., Hillyard, S.A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. Science 207, 203-205.

Kutas, M., Hillyard, S.A. (1984). Brain potentials during reading reflect word expectancy and semantic association. Nature 307, 161-163.

Litcofsky, K.A., Van Hell, J.G. (2017). Switching direction affects switching costs: Behavioral, ERP and time-frequency analyses of intra-sentential codeswitching. Neuropsychologia 97, 112-139.

Martin, A. (2007). The representation of object concepts in the brain. Annual Review of Psychology 58, 25-45.

Matsumoto, A., Iidaka, T., Haneda, K., Okada, T., Sadato, N. (2005). Linking semantic priming effect in functional MRI and event-related potentials. Neuroimage 24, 624-634.

Meulman, N., Stowe, L.A., Sprenger, S.A., Bresser, M., Schmid, M.S. (2014). An ERP study on L2 syntax processing: When do learners fail? Frontiers in Psychology 5, art, 1072, 17p.

Monsell, S., Doyle, M.C., Haggard, P.N. (1989). Effects of frequency on visual word recognition tasks - Where are they? Journal of Experimental Psychology: General 118, 43-71.

Moore, B.C.J., Creeke, S., Glasberg, B.R., Stone, M.A., Sek, A. (2012). A version of the TEN Test for use with ER-3A insert earphones. Ear and Hearing 33, 554-557.

Ohala, J.J. (1988). "The phonetics and phonology of aspects of assimilation," in *Papers in Laboratory Phonology I – Between the Grammar and Physics of Speech*, edited by Kingston, J., Beckman, M.E. (Cambridge University Press, Cambridge, United Kingdom), 258-275.

Oostenveld, R., Fries, P., Maris, E., Schoffelen, J.M. (2011). FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. Computational Intelligence and Neuroscience 2011, art. 156869, 9 p.

Osterhout, L., Nicol, J. (1999). On the distinctiveness, independence, and time course of the brain responses to syntactic and semantic anomalies. Language and Cognitive Processes 14, 283-317.

Osterhout, L., Holcomb, P.J. (1992). Event-related brain potentials elicited by syntactic anomaly. Journal of Memory and Language 31, 785-806.

Proverbio, A.M., Riva, F., Zani, A. (2010). When neurons do not mirror the agent's intentions: sex differences in neural coding of goal-directed actions. Neuropsychologia 48, 1454-1463.

Romei, L., Wambacq, I.J.A., Besing, J., Koehnke, J., Jerger, J. (2011). Neural indices of spoken word processing in background multi-talker babble. International Journal of Audiology 50, 321-333.

Steffensen, S.C., Ohran, A.J., Shipp, D.N., Hales, K., Stobbs, S.H., Fleming, D.E. (2008). Gender-selective effects of the P300 and N400 components of the visual evoked potential. Vision Research 48, 917-925.

Steinhauer, K., Abada, S.H., Pauker, E., Itzhak, I., Baum, S.R. (2010). Prosody-syntax interactions in aging: Event-related potentials reveal dissociations between on-line and off-line measures. Neuroscience Letters 472, 133-138.

Swaab, T.Y., Ledoux, K., Camblin, C.C., Boudewyn, M.A. (2011). Language-related ERP components. In: Kappenman, E.S., Luck, S.J. (Eds.), The Oxford Handbook of Event-Related Potential Components. Oxford Handbooks Online, Oxford, United Kingdom, 49 p.

Thornton, A.R.D. (2007). Instrumentation and recording parameters. In: Burkard, R., Don, M., Eggermont, J. (Eds.), Auditory Evoked Potentials: Basic Principles and Clinical Application. Lippincott William & Wilkins, Baltimore, MD, pp. 73-101.

Titone, D.A., Salisbury, D.F. (2004). Contextual modulation of N400 amplitude to lexically ambiguous words. Brain and Cognition 55, 470-478.

Tsolaki, A., Kosmidou, V., Hadjileontiadis, L., Kompatsiaris, I.Y., Tsolaki, M. (2015). Brain source localization of MMN, P300 and N400: Aging and gender differences. Brain Research 1603, 32-49.

Van Heuven, W.J.B., Mandera, P., Keuleers, E., Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. The Quarterly Journal of Experimental Psychology 67, 1176-1190.

Van Petten, C., Luka, B.J. (2006). Neural localization of semantic context effects in electromagnetic and hemodynamic structures. Brain and Language 97, 279-293.

Van Petten, C., Luka, B.J. (2012). Prediction during language comprehension: benefits, costs, and ERP components. International Journal of Psychophysiology 83, 176-190.

Valderrama, J.T., de la Torre, A., Van Dun, B. (2018). An automatic algorithm for blink-artifact suppression based on iterative template matching: application to single-channel recording of cortical auditory evoked potentials. Journal of Neural Engineering 15, art. 016008, 16 p.

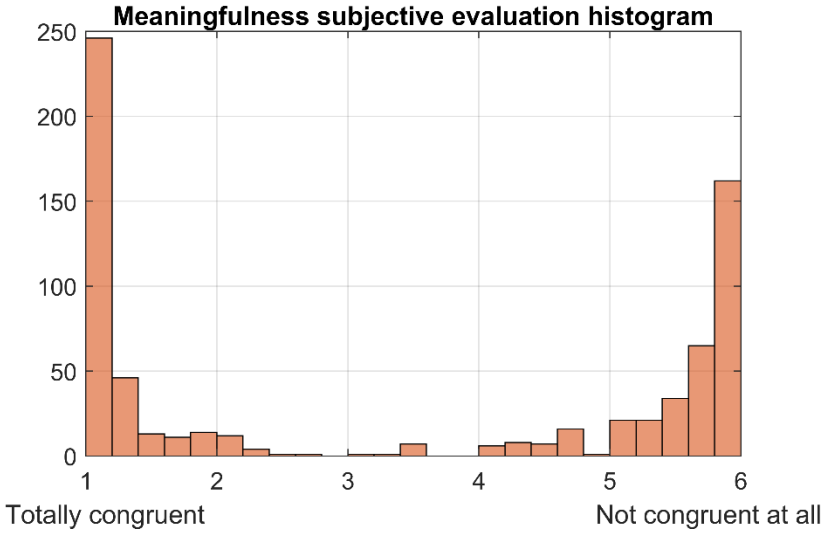
Wilkinson, A.J., Yang, L., Dyson, B.J. (2013). Modulating younger and older adult's performance in ignoring pictorial information during a word matching task. Brain and Cognition 83, 351-359.

Wirth, M., Horn, H., König, T., Stein, M., Federspiel, A., Meier, B., Michel, C.M., Strik, W. (2007). Sex differences in semantic processing event-related brain potentials distinguish between lower and higher order semantic analysis during word reading. Cerebral Cortex 17, 1987-1997.

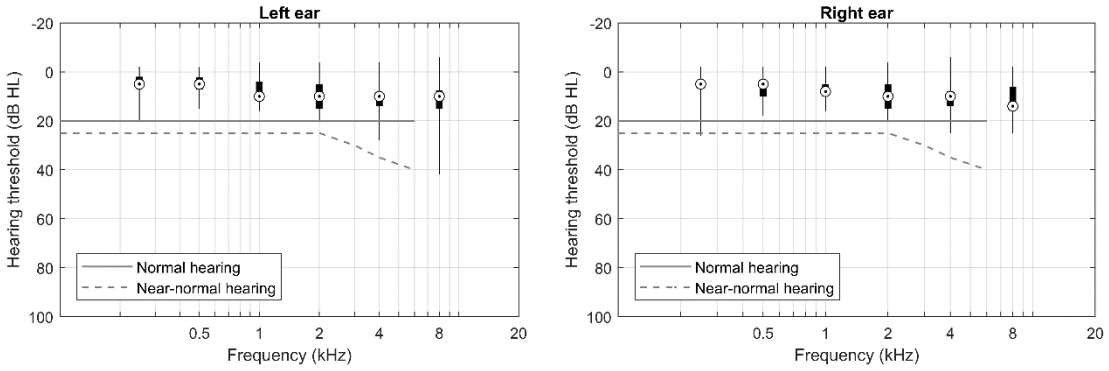
Zipf, G.K. (1949). Human behaviour and the principle of least effort. Oxford, England: Addison-Wesley Press, 573 p.

Figures

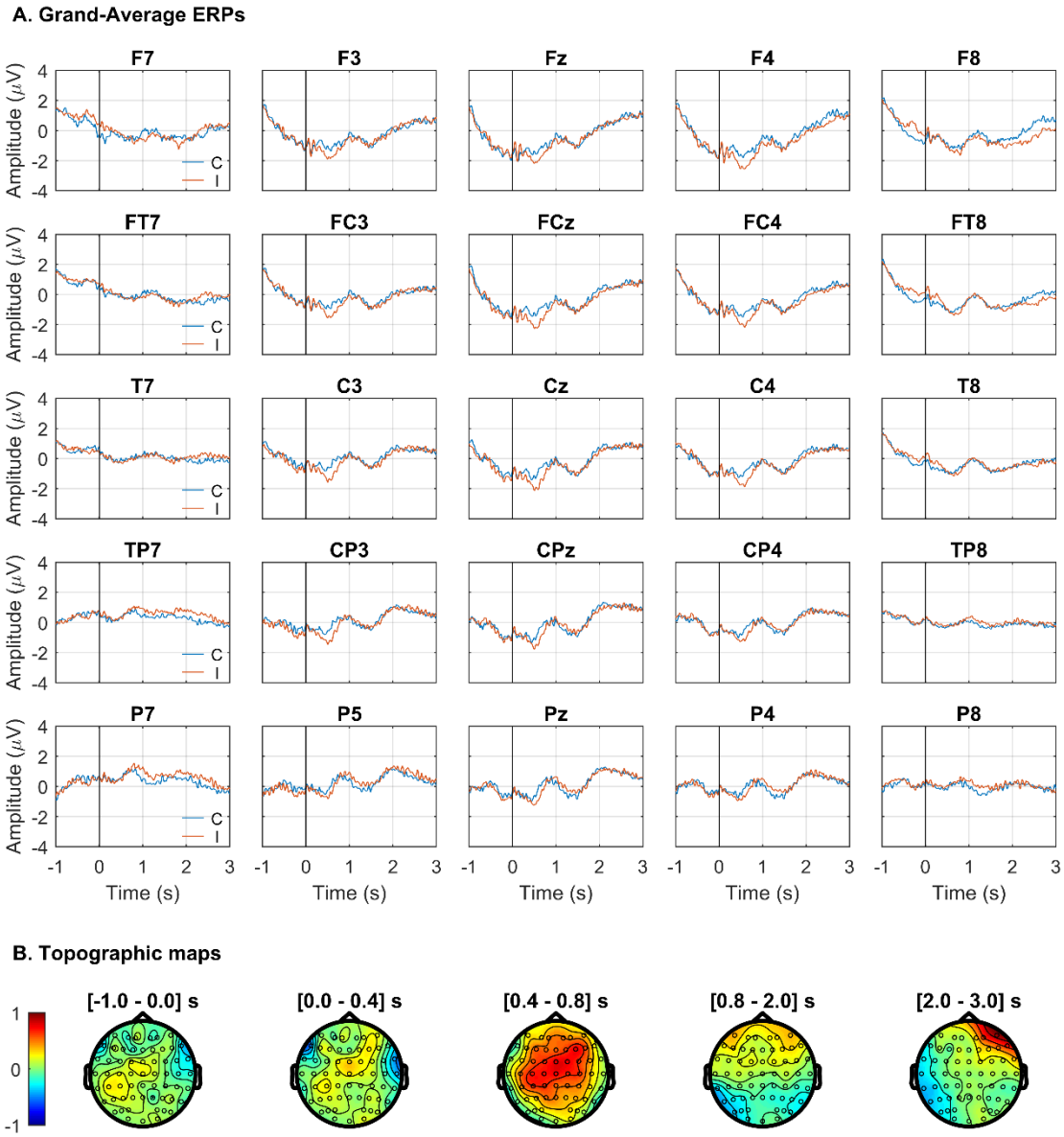
- Figure 1. Histogram of the subjective meaningfulness ratings for the list of congruent and incongruent sentences.



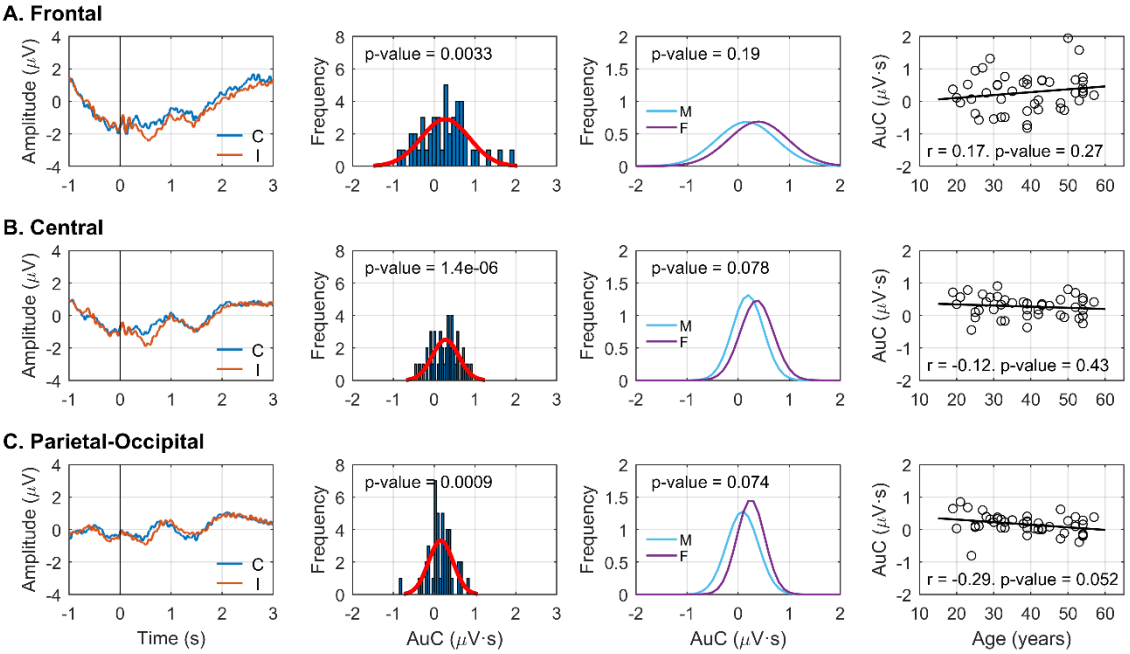
- Figure 2. Pure-tone audiometric threshold distributions for left and right ears. Boxplots indicate the minimum, 1st quartile, median, 3rd quartile, and maximum values of the distributions. Straight and dashed lines indicate the limits for normal and near-normal hearing, respectively.



- Figure 3. [A] Grand-Average N400 ERPs for selected EEG channels. [B] Topographic plots show the scalp distribution of the area under the curve estimated from the Grand-Average ERPs at different time intervals.



- Figure 4. Cluster analysis for the area under the curve (AuC) in frontal, central, and parietal-occipital areas: [1st column] Grand-average cluster ERPs; [2nd column] raw AuC distribution, fitted distribution, and probability (p) that the distribution comes from a normal distribution with mean equal to 0; [3rd column] fitted distribution for males and females, p-value indicating the probability that the two distributions present the same mean; [4th column] effect of age on AuC evaluated through a linear regression analysis.



Tables

- Table 1. Example of the three types of sentences. Sentences 1a and 1c are congruent sentences that were recorded naturally. Sentence 1b is an incongruent sentence artificially constructed from sentences 1a and 1c by cross-splicing the critical word. The critical words in these sentences are highlighted in bold.

1a	The driver puts the **petrol** in the car
1b	The mother breaks the **petrol** on the shelf
1c	The mother breaks the **glasses** on the shelf