The Sound of Learning: Prosodic Indicators of Memory Performance and Subjective Confidence in L1 and L2

Edith Fernández Amurrio

s5059852

Department of Psychology, University of Groningen

PSB3E-BT15: Bachelor Thesis

Group 19

Supervisor: dr. A. Sarampalis

Second evaluator: dr. O. C. Dimigen

In collaboration with: Y. B. Daskalova, A. A. Jarczynska, A. F. H. J. Lukassen,

T. Peters, E. E. C. Velthuizen

June 27th, 2025

A thesis is an aptitude test for students. The approval of the thesis is proof that the student has sufficient research and reporting skills to graduate, but does not guarantee the quality of the research and the results of the research as such, and the thesis is therefore not necessarily suitable to be used as an academic source to refer to. If you would like to know more about the research discussed in this thesis and any publications based on it, to which you could refer, please contact the supervisor mentioned.

Abstract

Previous research has shown that prosodic speech features distinctively reflect one’s accuracy and subjective confidence on an answer. These features can be used to enhance the effectiveness of adaptive learning systems and thus help provide a more personalised learning experience. The current study investigated how prosodic features of sentences (speaking speed, intensity and pitch) reflect accuracy and subjective confidence in individuals learning a language. We also assessed the strength of these relationships in native language utterances compared to second language utterances. In this experiment, native Dutch speakers studied and verbally retrieved short subject-verb sentences in Dutch and Italian in a counterbalanced order. Our results indicate that accuracy in an answer was best reflected by a higher mean pitch and lower mean intensity, whereas confidence was best reflected by a higher speaking speed and a larger pitch fall at the end of the sentence. Additionally, prosodic features were more informative of one’s learning performance in the native language than in the second language. Our findings indicate that prosodic information, especially in the native language, could be a valuable tool to improve adaptive learning systems’ estimations of a learner’s knowledge and feeling of knowing. We recommend that future researchers focus on applying these findings to real-world language learning contexts.

Keywords: PSFs, accuracy, confidence, language, ALSs

The Sound of Learning: Prosodic Indicators of Memory Performance and Subjective Confidence in L1 and L2

In daily conversation, individuals combine several linguistic elements, such as grammar, phonetics and morphology to convey messages. However, there is an additional element that contributes to the meaning of utterances, known as speech prosody. Prosody is characterized by changes in intonation, speaking speed and intensity, among others (Prieto & Roseano, 2018; Reed, 2011; van Maastricht, 2018). We will refer to these suprasegmentals – i.e., sound properties beyond the phonemes – as prosodic speech features (PSFs). Intonation refers to the pitch or absolute frequency of an utterance. This melodic feature often serves to indicate the modality of a sentence (i.e., affirmation, question, imperative) and the subject's epistemic stance on their spoken message (Roseano et al., 2015). For instance, the sentence 'Oh nice, another meeting' can be understood literally or ironically depending on changes in pitch and speaking speed across the utterance (van Maastricht, 2018). Speaking speed can be defined as the number of syllables per second in a speech signal. Intensity describes, in simple terms, the loudness of (parts of) an utterance. In some languages, changes in intensity are indicators of lexical stress (Koffi, 2019).

While PSFs can be used to willingly emphasize or clarify a message, these features also carry information about the inner states of the speaker, such as one's knowledge and their feeling of knowing. PSFs have been found to reflect objective accuracy and subjective confidence on an answer in several experimental contexts (Goupil & Aucouturier, 2021; Jiang & Pell, 2017), including language learning tasks (Wilschut et al., 2023; Wilschut et al., in review). In this study, we will expand on those findings and investigate how PSFs reflect the accuracy and confidence of individuals learning a new language. This research is necessary to understand how we monitor and (implicitly) communicate our performance during learning. Along with its theoretical relevance, it has practical applications for adaptive learning systems

(ALSs). Knowing how PSFs indicate accuracy and, especially, confidence in a spoken answer may contribute to the improvement of cognitive models of learning, which are at the base of ALSs.

ALSs aim to make learning of facts and vocabulary more efficient in terms of time and effort by tailoring elements of the study session to the individual needs of the learner. These systems are commonly used for obtaining factual knowledge, such as definitions or second language (L2) vocabulary (Sense et al., 2021). In an adaptive learning study session, different elements can vary depending on the system – the schedule of item presentation, the difficulty of the items, the presence of feedback, etc (Papousek et al., 2014; Sense et al., 2021).

Several findings that informed the purpose and design of this study were obtained through experiments conducted using the Memorylab ALS. Memorylab uses an algorithm that optimizes learning by altering the presentation schedule of study items based on behavioral measures of performance. Knowledge retention is maximized by balancing the benefits of the spacing effect and the testing effect (Sense et al., 2019; van Rijn et al., 2009). This system is based on the ACT-R declarative memory model, which uses the memory activation of a study item to predict when the item will be forgotten (Anderson, 1998; Pavlik & Anderson, 2008). Because memory activation is a latent state of the learner and therefore not suitable for direct measurement, it is estimated from two behavioral measures: reaction time (RT) and accuracy. Then, the memory activation estimate for a response is used to calculate the rate of forgetting for that study item and to present the item shortly before it is predicted to be forgotten (Sense et al., 2016). ALSs have been shown to yield better learning outcomes than flashcard learning and traditional memorization, since they account better for the fact that learners differ in the ease with which they approach the study material (Mettler et al., 2016; Sense et al., 2021).

Although Memorylab was developed as a typing-based ALS – where accuracy scores are based on the discrepancy between the typed and the model answer and RTs are counted starting from the first keypress –, it has been successfully adapted for spoken input, yielding the same learning benefits. In a study by Wilschut et al. (2021), a speech-based adaptive learning system was compared to the typing-based system and to a flashcard learning condition. Native Dutch and German speakers learned English words with complex phonology and orthography. Participants completed three study sessions differing in modality (spoken or typed input) and scheduling (adaptive or flashcard), each followed by a test where they provided the translation of several English words. For spoken input, RTs were triggered from the onset of sound and accuracy scores were given manually by a researcher. Wilschut et al. (2021) found no differences in average accuracy between typing and speech-based adaptive scheduling conditions, and RTs were good predictors of memory performance in both conditions. Furthermore, the speech-based adaptive scheduling condition yielded, on average, shorter RTs and higher accuracy scores than the speech-based flashcard scheduling condition. Thus, learners benefit from speech-based ALSs at least as much as they benefit from typing-based ALSs.

More importantly, there are several advantages to speech-based ALSs. Firstly, it allows people that have difficulty typing due to a general impairment – e.g., dyslexia (Wilschut et al., 2025) – or a momentary circumstance, like cooking or driving, to benefit from an individualized learning approach. Secondly, speech and pronunciation practice becomes available to learners of an L2. This aspect is essential for attaining proficiency in a language, but it is overlooked in the typing-based Memorylab system. Lastly, and most importantly, spoken answers may provide the opportunity to refine ALSs, since they contain PSFs that reflect how confident learners feel about responses and how objectively accurate these responses are (Goupil & Aucouturier, 2021; Jiang & Pell, 2017).

Jiang and Pell (2017) performed an analysis of pitch, intensity and speed in spoken statements explicitly conveying different levels of confidence to determine how prosodic patterns reflect certainty. They found confident utterances to be faster, higher in mean intensity and lower in mean pitch. When zooming into local variation in a sentence, a falling pitch denoted a high confidence level, whereas a rising pitch denoted low confidence. Similar results were obtained by Goupil and Acouturier (2021), who examined whether subjective confidence and objective accuracy are associated with distinct prosodic features. In their visual detection task, participants had to verbally choose the target word between two similar ones and then report their confidence on their answer. More confident responses tended to have a rising-falling intonation pattern and a higher intensity, although, contrary to Jiang and Pell (2017)'s findings, the duration of the utterance was longer. In addition, subjective confidence was a significant predictor of the duration and intonation of an utterance, whereas accuracy significantly predicted the intensity of the utterance.

PSFs have also been found to be distinctively indicative of accuracy and confidence in language learning tasks using the speech-based Memorylab ALS, where participants learned the English translations of Swahili (Wilschut et al., 2023) and Lithuanian (Wilschut et al., in review) vocabulary words. In these tasks, accuracy – scored with automatic speech recognition – was better correlated with average intensity, and confidence was better correlated with pitch slope and speaking speed. That is, correct answers tended to be uttered louder and utterances with a rising intonation and a faster speech rate were more often unconfident ones. As previously described, the cognitive model behind the Memorylab system estimates memory activation from RTs and accuracy scores of previous answers to an item. In their study, Wilschut et al. (2023) found that predictions of future memory performance (in terms of response latency) improved when adding PSFs of previous item repetitions to the model. They elaborated on this idea in later research, suggesting that a model including pitch slope and speaking speed as predictors of metacognitive state (confidence) and intensity as a predictor of memory activation (accuracy and RT) accounts for more variance in the data than models where PSFs predict one single inner state (Wilschut et al., in review).

The purpose of this study is to better describe the relationship between PSFs and measures of learning performance (i.e., accuracy and confidence) by investigating these variables in an experimental task somewhat different to that of Wilschut et al. (in review). Instead of learning English translations of single words, participants will study and verbally recall pairs of simple subject-verb sentences in their native language (L1) and L2. Ultimately, if a learner wants to acquire proficiency in a language, they need to practice sentences – not just individual words – and focus not only on translating, but also on producing speech with correct pronunciation, which can only be achieved by speaking L2. With these changes in mind, we have two research aims.

Firstly, we want to investigate the extent to which accuracy and confidence are reflected in prosodic features of short sentences in L1. We will attend to L1 utterances to investigate this aim, since previous studies used a language of retrieval in which participants were proficient. Considering the similarities between our experimental task and that of Wilschut and colleagues (in review), we hypothesize similar correlations between the variables of interest. That is, we expect louder answers to be more indicative of objective knowledge and answers with a falling intonation and a higher speech rate – aligning with Jiang and Pell (2017) – to be more indicative of one's feeling of knowing. Moreover, our choice of using short sentences instead of words is partly grounded on the fact that sentences contain, by definition, more elements and thus, more prosodic information and variation than shorter utterances. Thus, PSFs of sentences may be more informative of accuracy and confidence on an answer than PSFs of words.

Secondly, since study items will be presented and retrieved in the participants' L1 as well as L2, we want to determine if PSFs in L1 and L2 differ from each other in how strongly they correlate with the confidence and accuracy of responses. We speculate that L1 PSFs features will be more reflective of learning performance measures than L2 PSFs. Due to the higher difficulty of the task in L2, participants may put most of their effort into performing well (i.e., recalling the correct answer), which could hinder *how* the answer is verbally produced.

Methods

Participants

We used a convenience sampling method, recruiting most participants through the SONA systems platform and some through our personal network. Both groups participated on a voluntary basis. Additionally, SONA participants (first-year BSc psychology students at the University of Groningen) were compensated with SONA points for course credit. We performed a power analysis for a significance level of .05 and a power of .80 based on the size of relevant correlations between PSFs and accuracy and confidence (r = -.12, r = .12 and r = -.18) found by Wilschut et al. (in review), and arrived at a desired sample size of 40. We invited a total of 50 participants to the study lab but, for unknown reasons, we could not retrieve experiment data from two participants. Thus, the final sample size was 48 (34 women, M = 19.9 years, SD = 1.6). All study participants met the following selection criteria: They were native Dutch speakers, had no speech or hearing impairments and had no previous formal knowledge of Italian (i.e., previous language lessons or experience with language learning applications that could affect task performance). These requirements were assessed in a background questionnaire, where demographic data (age and gender) was also collected. The design of this study was approved by the Ethics Committee of the University of Groningen with study code PSY-2223-S-0257, and all participants gave their informed consent before the experiment took place.

Materials

This experiment was programmed in JavaScript and HTML-5 using the jsPsych library (De Leeuw, 2015) and ran online using the JATOS platform (Lange et al., 2015). Participants completed the experiment in a quiet laboratory setting, in individual test cubicles. They carried out the tasks on a desktop computer and were given USB headphones with a microphone (the model was Nedis Xyawyon GHST100BK). We used the Google Web Speech API to transcribe spoken answers in real time.

For the study tasks, we created 40 simple subject-verb sentences (Table A1) with their respective Dutch (L1) and Italian (L2) translations and generated spoken stimuli for these sentences using Narakeet (n.d.). All sentences were unique (i.e., no nouns or verbs were repeated across sentences) and followed a subject-verb structure in the simple present tense. To reduce systematic variability in the stimuli that could influence the difficulty of the task, none of the words in a sentence shared the same root between the Italian and Dutch translations. For the same reason, we consistently used logical sentences (i.e., sensical subject-verb combinations, such as 'The tree blooms') and avoided the use of reflexive verbs. Out of the 40 sentences, we arbitrarily selected four for the practice trials and 30 for the experimental trials (15 for each language condition).

In order to measure participants' subjective confidence in their answers, we included the following question in each trial: *'How confident are you about this answer?'*. The answer options were '*Not confident'*, '*A little confident'*, *'Quite confident'* and '*Very confident'*.

Design

This study used a within-subjects design with one independent variable and several dependent variables. The independent variable was the language of retrieval, with two conditions: Dutch (L1) and Italian (L2). These languages differ in their phonology, syntax and lexicon due to their historical roots. Hence, choosing Italian as L2 allowed us to prevent the influence of language familiarity on the study results. The order of language conditions, as well as the set of 15 sentences to learn in each condition, were assigned in a counterbalanced manner across participants. The sentences in each set were presented in a randomized order for each participant.

The dependent variables of this study were accuracy, RT (in milliseconds), subjective confidence and several PSFs: speaking speed, in syllables per second; mean intensity, in decibels (dB) Sound Pressure Level; mean pitch and pitch change, both measured in hertz (Hz). Measures of pitch and intensity were obtained through acoustic analysis in Praat 6.2.07 (Boersma, 2007). RTs were triggered from the onset of speech and subjective confidence was operationalized through the scale presented above.

Procedure

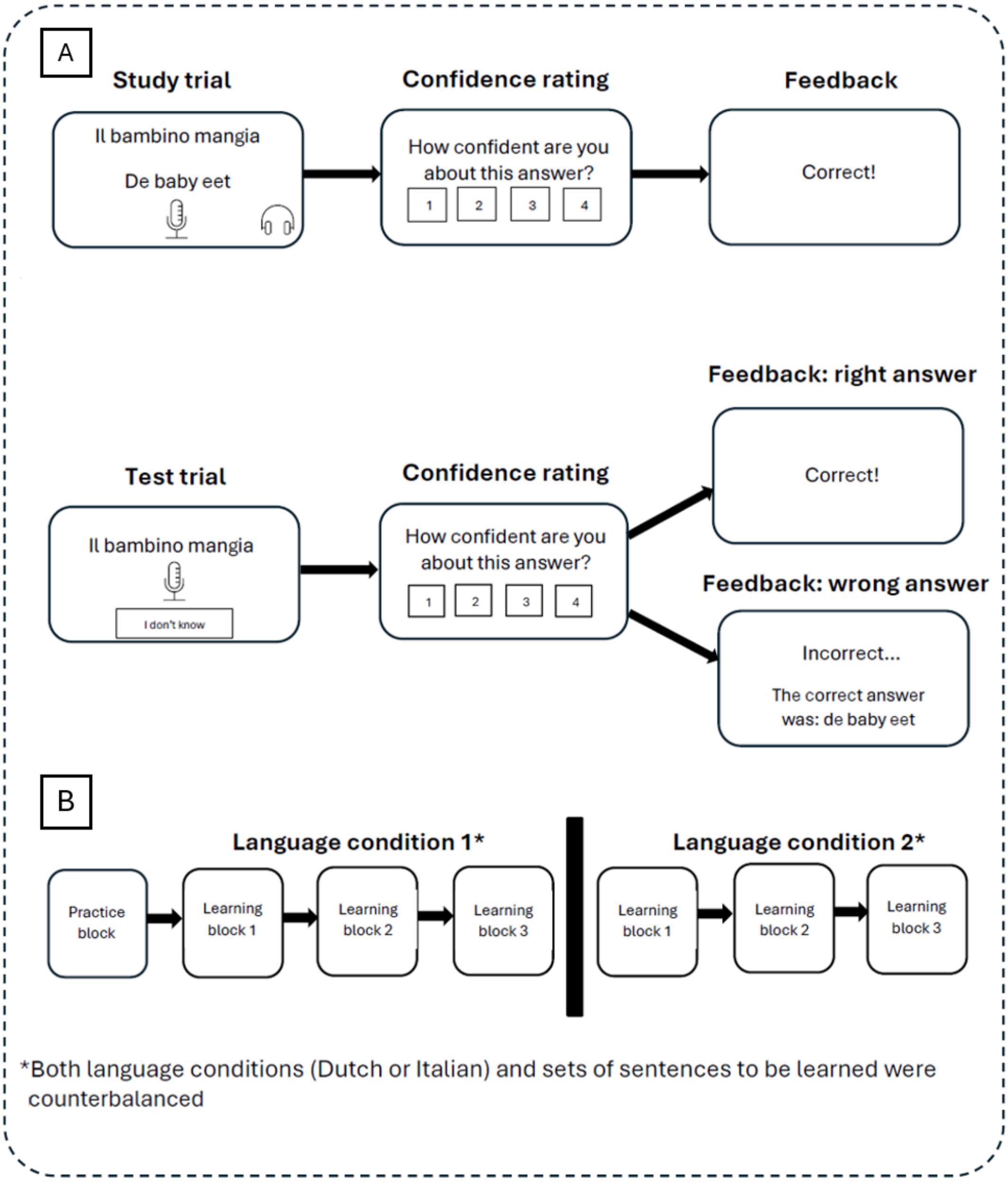
This experiment lasted approximately 30 minutes. Participants were first invited in the laboratory and assigned to an individual booth. Then, they read information about the study and gave their informed consent on a paper sheet in order to participate. They also completed a background questionnaire to confirm the absence of speech or hearing impairments and the lack of formal knowledge of Italian. Additionally, we asked participants to report their age, gender and any formal knowledge of other languages. After the questionnaire, we showed them the experiment on a desktop computer.

The language learning task consisted of studying and recalling short sentences through a series of study trials and test trials. Trials – both practice and experimental – were structured in the manner shown in Figure 1 (panel A): First, a sentence in Dutch or Italian was presented in the upper part of the screen. At the same time, its translation in the other language appeared in the lower part of the screen and was played via headphones. This is the item participants would be asked to retrieve later on. A speaker icon was also shown in the lower part of the screen to indicate that utterances would be recorded. After reading and hearing the stimuli, participants spoke out loud the translation of the sentence, which was recorded through the headset microphone, transcribed by the Google Web Speech API in real time, and stored for later processing and analysis. If participants could not remember the answer, they had the opportunity to click on an ‘I don’t know’ box. Next, they indicated certainty about their answer by selecting one of the options in the aforementioned confidence scale. Then, participants received written feedback on their answer. Feedback was automatically elaborated based on a comparison between the model answer (read and heard by the participant in the study trial) and the answer transcribed by the API. Answers were marked ‘correct’ if the transcription was at a Levenshtein’s edit distance of two or less from the model answer. When an answer was marked as incorrect, the participant received feedback with the correct response. It is worth noting that study trials and test trials were almost identical. The only difference between them was the presence or absence of the item to be retrieved. In study trials, participants were getting acquainted with a sentence for the first time and they could see its translation, whereas in test trials, only the sentence was shown and participants were expected to verbally recall the translation.

Participants completed one practice block before diving into the learning task (see Figure 1B). This block entailed studying four sentences and retrieving each of them twice in the language of the first experimental condition, making a total of 12 trials. Then, they carried out the main experimental task. This consisted of six blocks in total, three for each condition, with an opportunity to take a short break between conditions. Each block entailed the practice of five new sentences following a fixed schedule of presentation. That is, all sentences were presented the same amount of times and retrieved with equal spacing from one another. Participants first studied all sentences once and then successively retrieved them four times. This amounted to 25 trials per block (five study trials and 20 test trials) and 150 experimental trials in total across all blocks. After completing the experimental task, we debriefed the research aims of this experiment to the participants that expressed interest.

Figure 1

Schematic Illustration of the Experimental Procedure



Note. Panel A shows the principal elements of study and test trials. The headphones icon in the ‘study trial’ box indicates that the sentence to be retrieved was played to the participant. In the ‘confidence rating’ boxes, numbers one to four are a simplified representation (and our operationalization) of the written answer options that participants actually saw. Panel B displays the design of our language learning task.

Results

Before performing any statistical analyses, we pre-processed the data in several ways. First, subjective confidence on an answer was numerically defined on a scale from 1 (*'Not confident at all'*) to 4 (*'Very confident'*). Additionally, we calculated the speaking speed of utterances by dividing the speech duration in seconds by the number of syllables in each sentence. Next, we obtained the acoustic data for intensity and pitch through speech analysis in Praat 6.2.07 (Boersma, 2007). This computer software automatically discards empty audio files and uses the remaining ones to obtain numerical data suitable for statistical analysis. Using Praat, we computed the average intensity of utterances (meaning the spoken answers in each trial) and computed utterance pitch in 20 consecutive segments of equal length, as well as overall mean pitch for each utterance. Furthermore, we used the segments to calculate a measure of pitch change as comparable as possible to the 'pitch slope' variable in the study by Wilschut and colleagues (in review). The change in pitch was calculated by subtracting the average pitch of the first five segments from the average pitch of the last five segments. Lastly, it is important to note that we could not obtain intensity and pitch data for eight out of 48 participants due to missing recordings.

On an exploratory basis and to discard a potential confound in the study results, we assessed the mean API confidence for each participant in study trials in both language conditions. That is, we looked at how 'confident' the Google API was about the correctness of the transcription of spoken answers. We investigated study trials because participants had access to the correct answer through audio and in text, meaning that utterances – and consequently, their transcriptions – should almost always match the model answer. We found that the average API confidence was lower in the Dutch condition (M = 0.80, SD = 0.11) than in the Italian condition (M = 0.96, SD = 0.02). For nine participants, the mean API confidence on Dutch transcriptions was lower than 0.75, due to several trials where the API had reported very low confidence levels. Nevertheless, we decided not to discard those trials, given that they were little compared to the number of trials where the API was confident of having transcribed the answer correctly. Although one would expect the API to have more difficulty transcribing L2 than L1 speech (Ashwell & Elam, 2017), the Italian transcriptions were more reliable than the Dutch ones, perhaps because the API might be better trained in the former than in the latter.

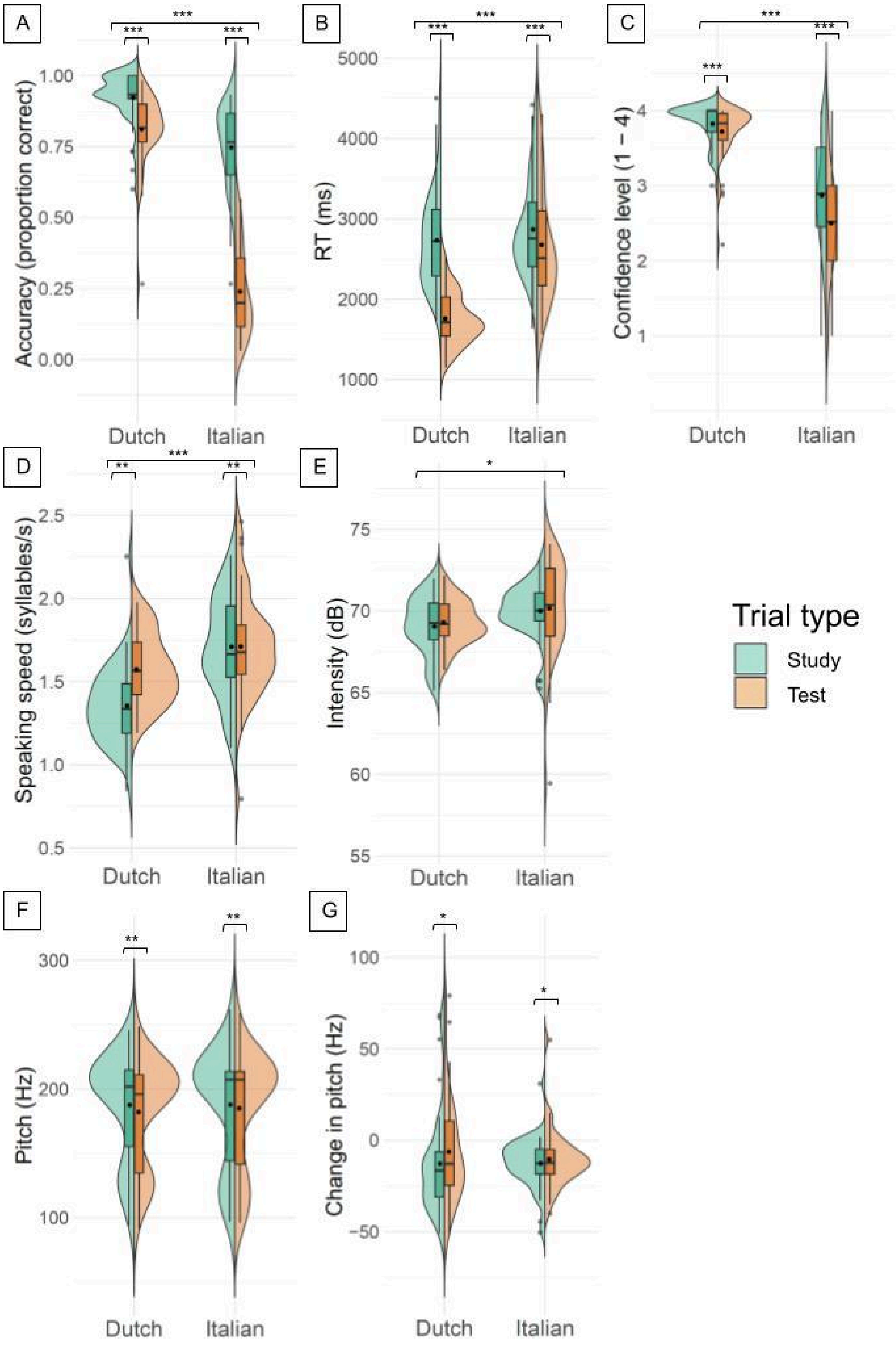
We performed all statistical analyses in JASP (Version 0.19.3; JASP Team, 2024) and used R (v.4.4.2, R Core Team, 2021) to visualize the data using several packages: *tidyverse* version 2.0.0 (Wickham et al., 2019), *introdataviz* version 0.0.0.9003 (Nordmann et al., 2022), *corrplot* version 0.95 (Wei & Simko, 2024) and *ggplot2* version 3.5.2 (Wickham, 2016).

Descriptive Analysis

Firstly, we computed means and standard deviations for all dependent variables on a participant level, as a function of language (Dutch, Italian) and trial type (study, test). These are plotted in side-by-side violin plots in Figure 2. We only included correct trials in this analysis, with the exception of accuracy descriptives, which were obtained from correct and incorrect trials. Furthermore, in order to describe and ascertain differences between groups in terms of condition and type of trial, we also performed seven repeated measures Analyses of Variance (RM ANOVA), one for each behavioral and acoustic dependent variable. In the model, the within-subjects independent variables were language condition (Dutch, Italian) and trial type (study, test), and the dependent variable was a different one in each RM ANOVA: accuracy, RT, subjective confidence, speaking speed, average intensity, average pitch and pitch change. The normality assumption of the RM ANOVA model was checked with a QQ plot for each variable. The assumption was met for all variables except for average pitch. We decided to still proceed with the analysis for that variable, since the RM ANOVA model is robust to violations of normality. Additionally, we were interested in finding out which specific groups differed from each other beyond the main effects and the interaction effect. Thus, we performed post-hoc comparisons with the Bonferroni correction. It is important to note that acoustic data were missing for nine participants, resulting in a smaller sample size and less degrees of freedom in the analyses for average intensity, average pitch and pitch change. The main results of all RM ANOVAs are summarized in Table 1.

Figure 2

Side-by-side Violin Plots for all Dependent Variables, Split by Language and Trial Type



Note. $*p < .05$,$**p < .01$, $***p < .001$. Accuracy was plotted using data from correct and incorrect trials. All other violin plots were created with data from correct trials only. The x-axis represents the language condition

(Dutch, Italian), whereas the y-axis represents the dependent variable, including its measurement units. Additionally, the green half of a plot indicates a study trial and the orange half indicates a test trial, thus creating four within-subjects groups per variable. All plots display kernel density estimates of the distribution of a group. Due to smoothing, violin shapes extend beyond the theoretical bounds for accuracy (0 to 1) and confidence (1 to 4). The figure shows the significant effects of language condition (square bracket encompassing two violin plots) and/or trial type (separate square brackets for each plot), not the differences between specific groups.

In terms of accuracy, study trials (*M* = 0.84, *SE* = 0.02) were significantly more correct than test trials (*M* = 0.53, *SE* = 0.02). Similarly, answers uttered in Dutch (*M* = 0.87, *SE* = 0.02) were significantly more accurate than answers in the Italian condition (*M* = 0.49, *SE* = 0.02). In addition, there was a significant interaction effect: the difference in accuracy between study and test trials was much larger in the Italian condition than in the Dutch condition. Post-hoc comparisons revealed significant differences between all the groups – all p's < .001 except for the Dutch test trial and Italian study trial comparison (*t* = 2.76, *p* = .042). Panel A displays these differences in the form of boxplots.

RTs (Panel B) were significantly slower in study trials (*M* = 2801.62, *SE* = 69.79) compared to test trials (*M* = 2217.19, *SE* = 69.79). At the same time, Dutch trials (*M* = 2246.06, *SE* = 66.37) had significantly faster reaction times than Italian trials (*M* = 2772.76, *SE* = 66.37). Additionally, there was a significant interaction effect, where the difference in RTs between study and test trials was larger for the Dutch condition than for the Italian condition. Lastly, post-hoc tests revealed significant differences only for three pairs: Dutch study trials compared to Dutch test trials (*t* = 9.50), Dutch test trials compared to Italian test trials (*t* = -9.85) and Dutch test trials compared to Italian study trials (*t* = -10.514, all *p*'s < .001).

Confidence ratings (Panel C) were higher in study trials (*M* = 3.35, *SE* = 0.06) compared to test trials (*M* = 3.11, *SE* = 0.07). In addition, participants reported more confidence when they retrieved sentences in Dutch (*M* = 3.77, *SE* = 0.04) compared to Italian (*M* = 2.69, *SE* = 0.09). There was a significant interaction effect. That is, participants felt more confident in study trials than in test trials, and this difference was more pronounced in the Italian condition. All post-hoc tests were significant (all *p*'s < .001) except for the test comparing Dutch study and test trials (*t* = 1.37, *p* = 1).

Regarding speaking speed (Panel D), utterances were significantly slower in study trials (*M* = 1.53, *SE* = 0.03) compared to test trials (*M* = 1.64, *SE* = 0.03). The model also found a significantly lower speech rate in the Dutch condition (*M* = 1.47, *SE* = 0.03) than in the Italian condition (*M* = 1.71, *SE* = 0.03). Lastly, there was a significant interaction effect, where the difference in speaking speed between study and test trials was larger when the answer was uttered in Dutch. Post-hocs comparisons revealed significant differences between all the groups (all p's < .01) except for the comparison between Italian study trials and Italian test trials (*t* = -0.03, *p* = 1).

Average intensity (Panel E) was lower (although not highly significant) in the Dutch condition (*M* = 69.29, *SE* = 0.26) compared to the Italian condition (*M* = 70.07, *SE* = 0.26). However, the main effect of trial type was nonsignificant, as well as the interaction effect. In line with these results, the post-hoc comparisons found no significant differences between the groups of means (all *p*'s > .05).

There was a significant effect of trial type on average pitch. Study trials (*M* = 188.30, *SE* = 6.72) had a slightly higher mean pitch than test trials (*M* = 184.40, *SE* = 6.72). Contrarily, there was no significant difference in average pitch between the two conditions, as well as no significant interaction effect (see Panel F). Additionally, only the post-hoc comparison between the Dutch study trials and test trials was significant (*t* = 2.77, *p* = .042), all others were nonsignificant (all *p*'s > .05).

Concerning pitch change (Panel G), the RM ANOVA found a slightly significant effect for trial type, meaning that study trials (*M* = -12.61, *SE* = 3.15) had a more falling pitch than test trials (*M* = -8,21, *SE* = 3.15). However, there was no significant effect for the language condition or the interaction between factors. This result was also reflected in the post-hoc tests, which were all nonsignificant (all *p*'s < .05).

Table 1

Repeated Measures ANOVAs Summary Table

Variable	Main effect of condition	Main effect of trial type	Interaction effect
Accuracy	F (1, 47) = 595.28, p < .001, η^2_p = .927	F (1, 47) = 315.37, p < .001, η^2_p = .870	F (1, 47) = 235.56, p < .001, η^2_p = .834
RT	F (1, 47) = 59.64, p < .001, η^2_p = .559	F (1, 47) = 52.41, p < .001, η^2_p = .527	F (1, 47) = 37.93, p < .001, η^2_p = .447
Confidence	F (1, 47) = 204.68, p < .001, η^2_p = .813	F (1, 47) = 18.1, p < .001, η^2_p = .277	F (1, 47) = 6.2, p = .020, η^2_p = .116
Speaking Speed	F (1, 47) = 64.09, p < .001, η^2_p = .577	F (1, 47) = 11.59, p = .001, η^2_p = .198	F (1, 47) = 16.76, p < .001, η^2_p = .263
Average Intensity	F (1, 38) = 5.09, p = .030, η^2_p = .118	F (1, 38) = 0.35, p = .557, η^2_p = .009	F (1, 38) = 0.01, p = .933, η^2_p = 0
Average Pitch	F (1, 38) = 0.00, p = .961, η^2_p = 0	F (1, 38) = 8.04, p = .007, η^2_p = .175	F (1, 38) = 0.93, p = .341, η^2_p = .024
Pitch Change	F (1, 38) = 0.38, p = .540, η^2_p = .010	F (1, 38) = 4.19, p = .048, η^2_p = .099	F (1, 38) = 1.41, p = .242, η^2_p = .036

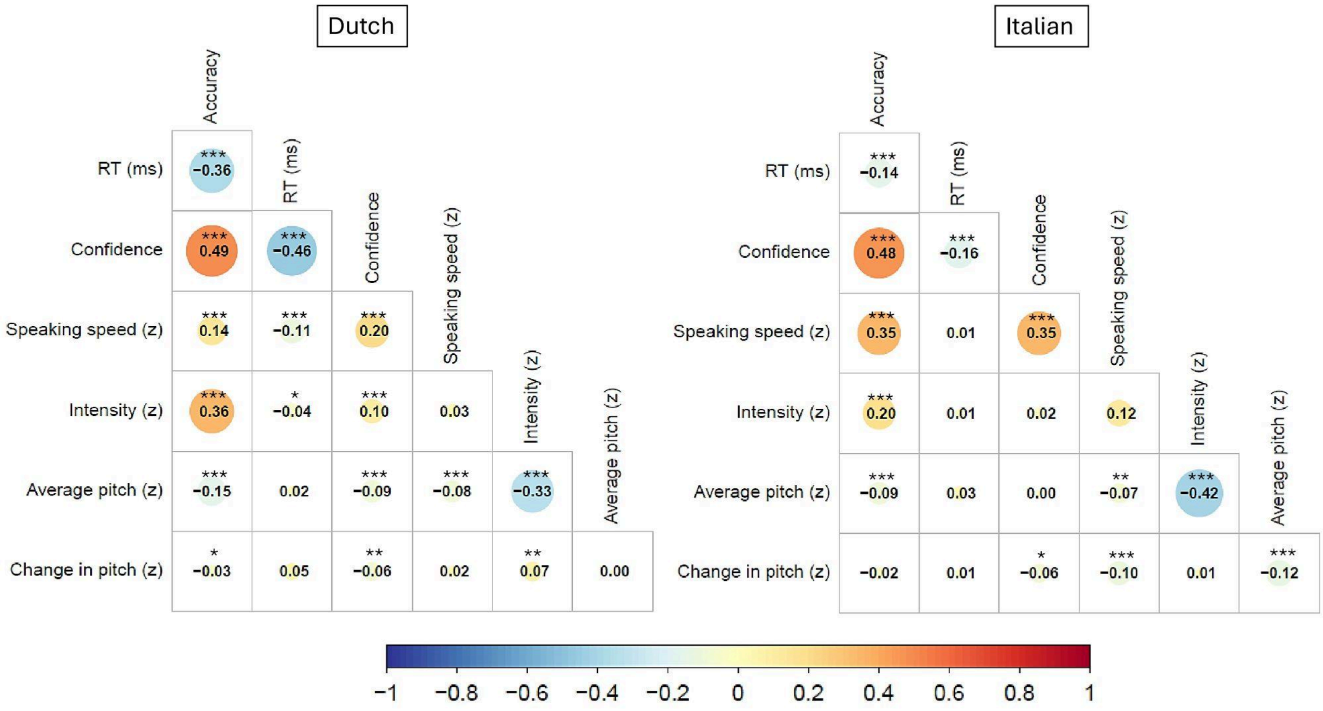
Note. Each cell displays, from left to right: F statistic with degrees of freedom, p-value and partial eta squared (as a measure of effect size). The RM ANOVAS were performed on averages of correct trials per participant, with the exception of the RM ANOVA for accuracy. The analysis for that variable was conducted with averages of both correct and incorrect trials.

Correlational Analysis

The first research aim of this project was to investigate the relationship between objective accuracy and subjective confidence on an answer, and PSFs in a task partly different to the one used by Wilschut et al. (in review). We were interested in finding out whether the correlations found in their study would be applicable to a language learning task that entails sentences instead of words. To this end, we computed Pearson's correlations between all behavioral and acoustic variables previously described, which are plotted in Figure 3. Acoustic variables were standardized (i.e. transformed into z-scores) to minimize the systematic variation between subjects. We correlated the variables on a trial level and only included test trials in the analysis, given that study trials do not require memory retrieval. More importantly, we performed the correlational analysis separately for the Dutch and Italian conditions. To investigate this research aim, we will attend to the correlations in the Dutch (L1) condition, since participants in the Wilschut et al. (in review) study were proficient in the language of item retrieval.

Figure 3

Correlation Matrices Split by Language Condition (Dutch, Italian)



Note. * *p* < .05, ** *p* < .01, *** *p* < .001 . Correlations were computed using data from test trials only. The correlation matrix displays behavioral variables first, followed by acoustics. RTs are measured in milliseconds, and all PSFs are standardized. In the figure, warmer colors represent more positive correlations and darker colors represent stronger correlations.

All behavioral measures were significantly correlated with each other. Accuracy was moderately correlated with RT (*r* = -.36, *p* < .001) and confidence (*r* = .49, *p* < .001), meaning that correct responses had faster reaction times and a higher reported confidence. We also found a medium sized negative correlation between RT and confidence (*r* = -.46, *p* < .001). Meanwhile, most correlations between behavioral and prosodic variables were significant but, as expected, not very strong. Speaking speed correlated the strongest with confidence (*r* = .20, *p* < .001), but in similar magnitude with accuracy (*r* = .14, *p* < .001) and RT (*r* = -.11, *p* < .001). In other words: Confident, accurate and fast responses tended to have a higher speech rate. Contrarily, mean intensity had a medium sized correlation with accuracy (*r* = .36, *p* < .001), and correlated weakly with RT (*r* = -.04, *p* = .039) and confidence (*r* = .10, *p* < .001). This indicates that correct answers were more often louder. Regarding the last two PSFs, average pitch had a small negative correlation with accuracy (*r* = -.15, *p* < .001) and confidence (*r* = -.09, *p* < .001), whereas pitch change had a very small correlation with RT (*r* = .05, *p* = .019) and confidence (*r* = -.06, *p* = .005). This indicates that utterances higher in pitch were slightly less correct and less confident, and that a more positive pitch at the end of an answer might be associated with slower RTs and less reported confidence. Lastly, only some of the PSFs correlated significantly with one another. Average pitch was negatively associated with speaking speed (*r* = -.08, *p* < .001) and mean intensity (*r* = -.33, *p* < .001), though the latter correlation was stronger. This indicates that answers with low mean pitch tended to be spoken louder and slightly faster. Similarly, pitch change correlated significantly with average intensity (*r* = .07, *p* = .002), implying that utterances ending with a rise in pitch were somewhat more intense.

The second aim of this study was to find out whether the association between markers of learning performance and PSFs in answers retrieved and spoken in L1 is different when the task is performed in L2. Hence, we now turn to the correlations in the Italian (L2) condition and compare them to those of the Dutch (L1) condition.

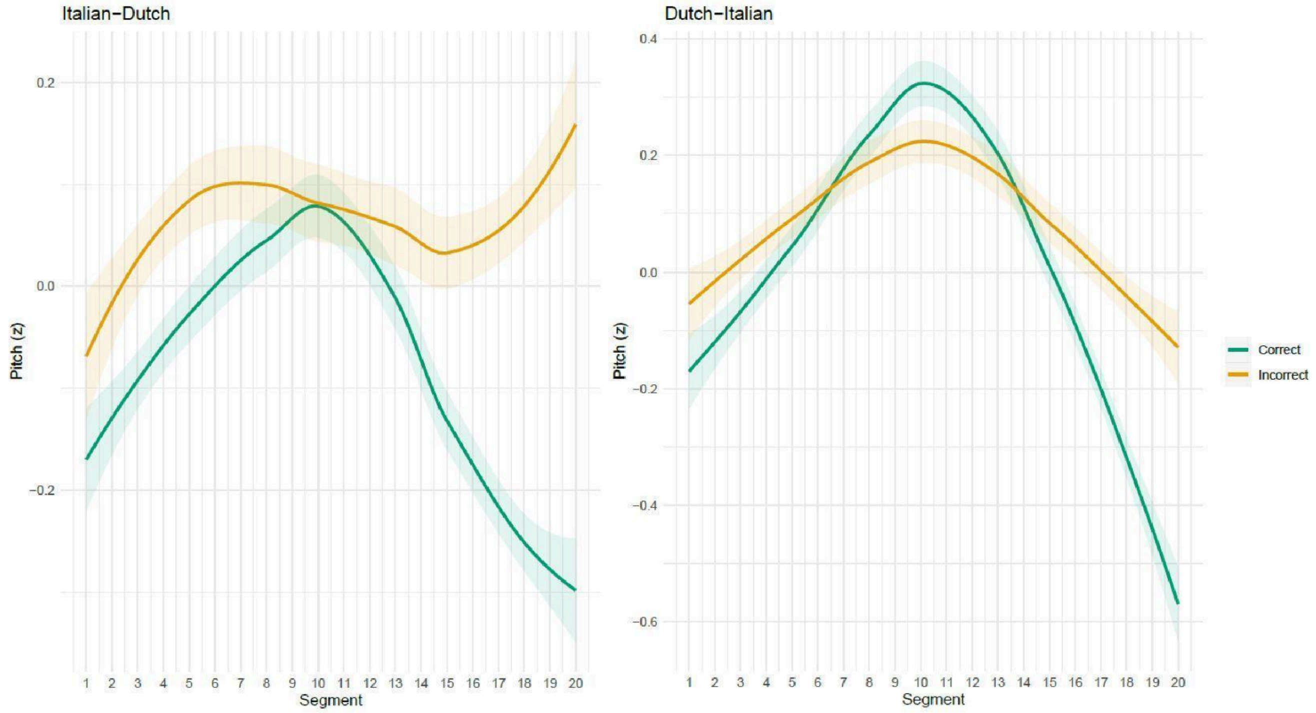
Like in the Dutch condition, accuracy was moderately correlated with confidence (*r* = .48, *p* < .001), meaning that participants were likely to be confident about correct answers. However, RTs were more weakly correlated with accuracy (*r* = -.14, *p* < .001) and confidence (*r* = -.16, *p* < .001) in Italian answers than in Dutch answers. Moreover, the correlations between the behavioral variables and the PSFs were generally less significant in the Italian condition compared to the Dutch condition. Speaking speed was an exception, and correlated significantly with confidence (*r* = .35, *p* < .001) and accuracy (*r* = .35, *p* < .001), with equal direction and strength. This indicates that utterances with a faster speech rate were more likely to be correct and have a high confidence rating, especially in the Italian condition. Meanwhile, the average intensity and pitch of utterances only correlated significantly with accuracy (r = .20, p < .001; r = -.09, p < .001), meaning that louder answers and answers with lower mean pitch tended to be somewhat more correct. In addition, a very small but significant correlation was found between pitch change and confidence (r = -.06, p = .014), indicating that a falling pitch was associated with higher confidence about the spoken answers. Lastly, more acoustic variables correlated significantly with one another in this condition than in the Dutch condition. For instance, speaking speed was significantly and positively correlated with average intensity (r = .12, p < .001), and negatively correlated with average pitch (r = -.07, p = .006) and change in pitch (r = -.10, p < .001), although the effect was small. This suggests that the higher the speech rate, the louder and lower in intensity the spoken answer was, also in association with a negative pitch change by the end of the sentence. Parallel to the Dutch condition, average pitch and average intensity were moderately and negatively associated with each other (r = -.42, p < .001). This negative relationship – the louder the utterance, the lower the intonation – was stronger in Italian utterances than in Dutch ones. Pitch change did not correlate significantly with intensity like in the Dutch condition. Instead, it was significantly correlated with average pitch (r = -.12, p < .001), implying that utterances higher in mean pitch were more likely to end with a decline in pitch.

Pitch Trajectory Analysis

As described earlier, we computed the change in pitch along an utterance as a subtraction, rather than a regression line following a trajectory across several points. This choice may have affected the significance and strength of the associations between pitch change and other variables, most importantly accuracy and confidence. To further investigate these relationships, we plotted the average trajectory of z-scored pitch across segments as a function of language condition, accuracy (Figure 4) and subjective confidence (Figure 5).

Figure 4

Average Standardized Pitch Trajectory as a Function of Language Condition and Accuracy

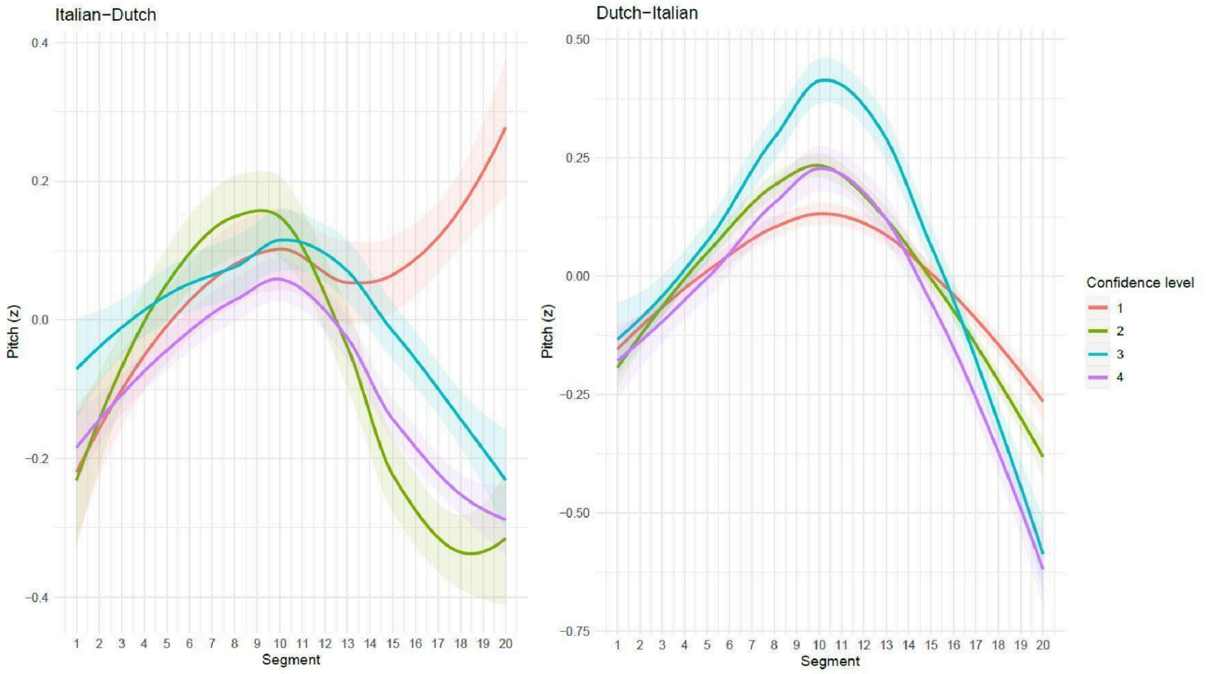


Note. The left panel displays data from the Dutch condition and the right panel displays data from the Italian condition. Only data from test trials was plotted. The x-axis is defined by the 20 segments in an utterance and the y-axis is defined by the z-scored pitch (minimum and maximum values differ in each panel). Smoothed lines represent averages calculated from test trials, where green indicates a correct answer – 81% of Dutch trials and 25% of Italian trials – and yellow indicates an incorrect answer – 19% of Dutch trials and 75% of Italian trials. The shaded areas represent 95% confidence intervals around the mean.

The average course of pitch in an utterance varied depending on whether the answer was correct or incorrect. In addition, this variation differed between the two language conditions. When Dutch utterances were correct, their pitch had a clear rising-falling pattern (see left panel). Contrarily, the pitch of incorrect answers rose in the beginning of the sentence, slightly declined in the middle segments and rose again in the end of the sentence. The mean pitch trajectory in Italian spoken answers followed a rising-falling pattern in both correct and incorrect answers, although the trajectory was steeper in correct answers, especially towards the end of the sentence (see right panel). When comparing the two conditions, pitch had larger and more abrupt changes across Italian utterances than across Dutch utterances.

Figure 5

Average Standardized Pitch Trajectory as a Function of Language Condition and Confidence



Note. The left panel displays data from the Dutch condition and the right panel displays data from the Italian condition. Only data from test trials was plotted. The x-axis is defined by the 20 segments in an utterance and the y-axis is defined by the z-scored pitch (minimum and maximum values differ in each panel). The smoothed lines represent averages calculated from test trials and the shaded areas represent 95% confidence intervals around the mean. Lines are split by confidence level on an answer (1 = ‘Not confident at all’, 2 = ‘A little confident’, 3 = ‘Quite confident’ and 4 = ‘Very confident’). Almost 75% of Dutch trials were scored with a 4, whereas 40% of Italian trials were scored with a 1, 30% with a 2 and 21% with a 3.

The level of confidence reported by participants influenced the (z-scored) pitch trajectory across utterances. In the Dutch condition, utterances with the lowest confidence level (where 1 stands for not feeling confident about the answer at all) ended with a rising pitch. However, all other subjective confidence levels had a clear rising-falling pitch trajectory (see left panel). Conversely, Italian utterances had a rising-falling intonation pattern in all confidence levels. Nevertheless, the fall in pitch in the last segments of an utterance was steeper in utterances with a higher confidence level (see right panel).

Discussion

In this study, we investigated the relationship between PSFs and markers of learning performance in a language learning task using short sentences spoken in L1 (Dutch) and L2 (Italian). Previous research has shown that speech prosody, as a high-level linguistic element, can reflect the objective accuracy and subjective confidence of a given answer. More specifically, correctness is better conveyed through higher speech intensity, whereas a higher speaking speed and a larger pitch drop at the end of an utterance reflect certainty on an answer (Goupil & Aucouturier, 2021; Jiang & Pell, 2017; Wilschut et al., in review). Such information about a learner’s memory strength and metacognition can be used to provide a more personalized learning experience (Wilschut et al., 2023). The present study built on these findings through two research aims: Firstly, we were interested in describing the extent to which accuracy and confidence are reflected by the PSFs of sentences. Additionally, we intended to make a comparison of the results between the L1 and the L2 condition.

Our study results show that participants were more correct and confident in study trials than in test trials and in L1 compared to L2 recall. In the Italian condition, participants could only recall the correct answer in 24% of the test trials. Given the high confidence of the API in Italian, we ruled out the possibility that low L2 accuracy scores could be due to transcription errors. Related to this point, Italian answers were considerably less confident than Dutch answers even in study trials. Interestingly, some of the participants reported that the task was easier in Dutch because they only needed to learn the association between an Italian cue sentence and its Dutch translation, whereas in the Italian task they additionally needed to recall the morphology and pronunciation of a sentence. Additionally, participants took more time to respond in study trials than test trials, especially in the Dutch condition. In the Italian condition, RTs decreased very little from study to test trials. We interpreted slower RTs in study trials as a mechanism for participants to process the new study items. In the Dutch condition, items were easier to learn, so it progressively took less time to remember them the more they were rehearsed. Contrarily, Italian answers required a similar amount of retrieving time because they were more difficult to remember. In sum, these results indicate that participants navigated the L1 task with more ease than the L2 task.

In connection with this, accuracy, confidence and RTs in test trials were associated with each other. In both language conditions, correct responses were rated with a higher confidence, meaning that participants engaged with the scale and were able to estimate their performance on the task quite accurately. Additionally, this finding suggests that the confidence scale we devised was a good operationalization of subjective confidence on an answer. Hence, we were able to reliably quantify subjective confidence, which was an important limitation in Wilschut et al. (in review)'s study – who used a slider that participants did not adequately engage with. Furthermore, RTs were negatively correlated with accuracy and confidence, especially in the Dutch condition. Considering the fact that RTs and accuracy scores are both indicators of memory activation strength, and that participants tended to be confident when they were correct, it follows that correct and confident answers had an earlier speech onset.

In response to our first research aim, we hypothesized a replication of Wilschut and colleagues' (in review) results in the L1 condition – except for speaking speed, where we expected findings similar to Jiang and Pell's (2017) –, perhaps with a larger effect size due to the nature of our study items. We found that accuracy and confidence were distinctly associated with PSFs in a manner similar to what had been found previously. Accuracy on an answer was best reflected by the average intensity and pitch of an utterance, whereas subjective confidence was best reflected by its speaking speed and change in pitch. More specifically, it is a common finding that louder responses are usually correct rather than incorrect (Goupil & Aucouturier, 2021; Wilschut et al., 2023; Wilschut et al., in review). What is more, we found a stronger correlation between the mean intensity of sentences and their accuracy than Wilschut et al. (in review) found with words, most likely due to the added prosodic information. Secondly, our results suggest that a lower pitch is more indicative of accuracy, although previous evidence suggested it is indicative of certainty (Jiang & Pell, 2017). Since Wilschut and colleagues (in review) did not examine average pitch and Jiang and Pell (2017) did not study the variable in the learning context, we can infer for now that a low mean pitch is more indicative of correctness in learning tasks. Several studies, including ours, have shown that a falling intonation reflects certainty, whereas a rising intonation reflects doubt (Jiang & Pell, 2017; Goupil & Aucouturier, 2021; Wilschut et al., in review). We found a weaker relationship between pitch change and subjective confidence than Wilschut et al. (in review), even though our study items contained more prosodic information. We attribute this to the fact that our estimation of local pitch variation was less nuanced than their pitch slope, largely ignoring any variation in the middle of the utterance. For this reason, we investigated the pitch trajectory of answers as a function of correctness and reported level of confidence. In L1, correct and confident answers had a rising-falling intonation, whereas incorrect and unconfident utterances ended with a rise in pitch. Lastly, a higher speaking speed denoted confidence on an answer, which is congruent with Jiang and Pell's (2017) findings, and contrary to the results of Goupil and Aucouturier (2021) and Wilschut et al. (in review). We believe this incongruence stems from the limited prosodic information contained in single words, compared to sentences.

For the second aim, we hypothesized stronger associations between PSFs and accuracy and confidence in L1 than L2. Indeed, that was largely the case. Average intensity, average pitch and pitch change correlated more significantly and/or strongly with accuracy and subjective confidence in L1 condition than in L2 condition. Speaking speed was the exception. Its correlations with accuracy and confidence were of equal strength and direction in Italian trials, and stronger than the associations in Dutch trials. Additionally, the pitch trajectory of L2 utterances was rising-falling for correct and confident, as well as incorrect and unconfident responses. That is, the cognitive and metacognitive state of the learner did not influence the intonation of L2 answers, although it did influence the intonation of L1 answers. There are several possible reasons for why PSFs reflect cognitive and metacognitive states better in L1. Firstly, it is possible that – considering the lack of fluency, proficiency and, perhaps, investment in the Italian language – most cognitive efforts went into performing the task at the most basic level (i.e., trying to recall the answer). The artificiality of this learning context may have affected the prosodic features of Italian answers. Secondly, perhaps participants were unfamiliar with Italian pronunciation and relied on the guidance offered by the spoken stimuli in study trials. In that case, they may have instinctively imitated the pronunciation, but also the prosody of the sample audio later on in test trials to enhance their performance. No emulation was necessary in the Dutch condition, given that native speakers know what their language 'sounds like'. Approaching the same task in a different manner can, of course, affect the outcome. Besides this, due to a mishap, about half of the Italian spoken stimuli had a different (male) voice from the rest of the stimuli (which had female voices), whereas the Dutch stimuli all had the same voice characteristics. This inconsistency likely contributed to participants' confusion and to their prosodic patterns, especially if they were imitating the acoustics of the audio.

We acknowledge that this study had limitations, and that these may have affected our findings and their generalizability, but could be solved in future studies. Firstly, it is problematic that the speech to text API was better trained in Italian than in Dutch because it gave rise to incorrect transcriptions and therefore, incorrect feedback to correct answers in some trials. This noise in data somewhat interfered with accuracy scores. As experimenters, we relied on this tool and did not develop it ourselves. We can merely suggest that future language learning studies using automatic speech recognition to score answers ought to choose a language of retrieval with a higher API confidence if they wish to avoid this limitation. Nevertheless, the proficiency of the API in the language of a study session is an important point to consider if speech-based ALS are to be launched in the future.

Secondly, participants found the Italian condition very difficult, which constrains the validity of some results. Most RM-ANOVAs were conducted with correct trials only, meaning that Italian means were calculated with considerably less data (more than 50% in test trials) than Dutch means. For a valid comparison between these two conditions, means should summarize a similar amount of data. To bridge this limitation, future studies could benefit from scaffolding. For example, participants could first learn a word and, later on, a short sentence that includes it. Another option could be to recruit participants with a beginners' level in L2. This solution has the additional benefit that learners might refrain from copying the prosody of the model answer, since they would be familiar with L2 pronunciation. Beyond that, sampling individuals that want to learn or are learning a language – and are therefore more motivated – would not only help assess if PSFs are more reflective of accuracy and confidence in this context, but also provide a sample that better represents potential system users.

Furthermore, our learning task had a fixed item presentation schedule, where all items were presented an equal amount of times and at equal temporal distance from each other. This design choice was made to avoid an imbalance in the amount of data among different study items or between language conditions. However, we would like to see if the results of this experiment would replicate in a study using an adaptive presentation schedule. This design would have more ecological validity, since adaptive scheduling is essential to the Memorylab ALS.

An additional idea for future research is to assess the informativeness of PSFs in a language learning task where the (possibly native) language of retrieval is not English or Dutch, but a language from a different family. If comparable results were found in languages with a different communicative use of intonation and rhythm, the robustness of the present phenomenon would be strengthened. Nevertheless, the usefulness of the current findings is compelling.

Conclusion

The present study added nuance to our knowledge of the relationship between PSFs and accuracy and confidence by replicating previous findings in a learning task with new characteristics. We found that intensity and overall pitch are more indicative of correctness, whereas speaking speed and pitch change are more indicative of certainty in a spoken sentence. We also found evidence that the prosody of answers in L1 is more informative of a learner's knowledge and feeling of knowing than the prosody of L2 answers. Currently, ALSs efficiently estimate memory strength from accuracy and RTs. Yet, there is no successful way to measure one's confidence on an answer other than asking about it. Consequently, although prosodic information may enhance estimations of memory activation, PSFs reflective of subjective confidence could become a proxy for learners' confidence in adaptive learning models. Items could be revised more frequently in a study session if the learner is not confident in their ability to remember or even pronounce them. In sum, this study shows how prosodic features, especially in L1, reflect cognition and metacognition in individuals learning a new language. These findings pave the way for the improvement of ALSs, which could enable individuals to practice a language in an even more personalized and efficient manner.

Acknowledgements

First of all, I am deeply thankful to Tassos Sarampalis for his extensive assistance, support and understanding throughout this project. Beyond that, he has fostered a comfortable and transparent working environment, where our learning process was always the priority. Additionally, I would like to thank Thomas Wilschut for all the methodological help, and for sharing part of his PhD project (including his manuscript) with us.

Furthermore, I thank Yordanka, Ola, Anne, Thijs and Noor, who have made this a very pleasurable semester. I am grateful for the opportunity to work in such a nice group and to go through all the frustration and excitement together. I give my special thanks to Noor, for the beautiful plots that she selflessly shared with all of us. I wish them best of luck in their next educational journey.

Lastly, I would like to thank my friends, my parents and my brother for all their unconditional support and confidence in me. I am also deeply thankful to my best friend and boyfriend Sam, who went through all the long hours and the madness with me, and trusted that I would do a good job. Most importantly, I am grateful to my grandmother, who I miss dearly, for giving me the opportunity to be here even when she didn’t want to see me go. Thanks for being proud of me no matter what.

References

Anderson, J. R., Bothell, D., Lebiere, C., and Matessa, M. (1998). An Integrated Theory of List Memory. *J. Mem. Lang. 38*(4), 341–380. https://doi.org/10.1006/jmla.1997.2553

Ashwell, T., & Elam, J. R. (2017). How accurately can the Google Web Speech API recognize and transcribe Japanese L2 English learners' oral production? *The JALT CALL Journal*, *13*(1), 59–76. https://doi.org/10.29140/jaltcall.v13n1.j212

Boersma, P. (2007). Praat: Doing phonetics by computer. http://www.praat.org/

De Leeuw, J. R. (2015). Jspsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior research methods, 47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Goupil, L., & Aucouturier, J. J. (2021). Distinct signatures of subjective confidence and objective accuracy in speech prosody. *Cognition*, *212*, 104661. https://doi.org/10.1016/j.cognition.2021.104661

JASP Team (2024). JASP (Version 0.19.3)[Computer software]

Jiang, X., & Pell, M. D. (2017). The sound of confidence and doubt. *Speech Communication, 88*, 106–126. https://doi.org/10.1016/j.specom.2017.01.011

Lange, K., Kühn, S. & Filevich, E. (2015). "Just Another Tool for Online Studies" (JATOS): An Easy Solution for Setup and Management of Web Servers Supporting Online Studies. *PLoS ONE 10*(7): e0134073. https://doi.org/10.1371/journal.pone.0134073

Mettler, E., Massey, C. M., & Kellman, P. J. (2016). A comparison of adaptive and fixed schedules of practice. *Journal of Experimental Psychology: General, 145*(7), 897–917. https://doi.org/10.1037/xge0000170

Narakeet [Text-to-speech tool]. (n.d.). *Narakeet.* https://www.narakeet.com

Nordmann, E., McAleer, P., Toivo, W., Paterson, H. & DeBruine, L. (2022). Data visualisation using R, for researchers who don't use R. *Advances in Methods and Practices in Psychological Science, 5*(2) https://doi.org/10.1177/25152459221074654

Pavlik, P. I., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied, 14*(2), 101–117. https://doi.org/10.1037/1076-898X.14.2.101

Papousek, J., Pelánek, R., & Stanislav, V. (2014). Adaptive practice of facts in domains with varied prior knowledge. *Educational Data Mining 2014*, 6–13.

Prieto, P., & Roseano, P. (2018). Prosody: Stress, Rhythm, and Intonation. In K. L. Geeslin (Ed.), The Cambridge Handbook of Spanish Linguistics (pp. 211–236). Cambridge: Cambridge University Press.

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org/ [Computer software]

Reed, B. S., (2011). *Analysing Conversation.* Macmillan Education UK. http://dx.doi.org/10.1007/978-1-137-04514-0

Roseano, P., González, M., Borràs-Comes, J., & Prieto, P. (2015). Communicating Epistemic Stance: How Speech and Gesture Patterns Reflect Epistemicity and Evidentiality. *Discourse Processes, 53*(3), 135–174. https://doi.org/10.1080/0163853X.2014.969137

Sense, F., Behrens, F., Meijer, R. R., & van Rijn, H. (2016). An Individual's Rate of Forgetting Is Stable Over Time but Differs Across Materials. *Topics in Cognitive Science, 8*(1), 305-321. https://doi.org/10.1111/tops.12183

Sense, F., Jastrzembski, T., Krusmark, M., Martinez, S., & van Rijn, H. (2019). An Integrated Trial-Level Performance Measure: Combining Accuracy and RT to Express Performance During Learning. *Proceedings of the 41st Annual Meeting of the*

Cognitive Science Society: Creativity + Cognition + Computation, CogSci 2019 (pp. 1029-1034). The Cognitive Science Society.

Sense, F., van der Velde, M., & van Rijn, H. (2021). Predicting University Students' Exam Performance Using a Model-Based Adaptive Fact-Learning System. *Journal of Learning Analytics*, *8*(3), 155-169. https://doi.org/10.18608/jla.2021.6590

van Maastricht, L. (2018). Second Language Prosody: Intonation and rhythm in production and perception. [Doctoral Thesis, Tilburg University]. https://research.tilburguniversity.edu/files/25772689/Van_Maastricht_Second_09_05_2018.pdf

van Rijn, H., van Maanen, L., & van Woudenberg, M. (2009). Passing the test: Improving Learning Gains by Balancing Spacing and Testing Effects. In A. Hoses, D. Peebles, & R. Cooper (Eds.), *Proceedings of the 9th International Conference on Cognitive Modeling* (pp. 108-114).

Wei,T., & Simko,V. (2024). *R package 'corrplot': Visualization of a Correlation Matrix*. (Version 0.95) [Computer Software], https://github.com/taiyun/corrplot.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

Wilschut, T., Sense, F., van der Velde, M., Fountas, Z., Maaß, S. C., & van Rijn, H. (2021). Benefits of Adaptive Learning Transfer From Typing-Based Learning to

Speech-Based Learning. *Frontiers in Artificial Intelligence*, *4*. https://doi.org/10.3389/frai.2021.780131

Wilschut, T., Sense, F., Scharenborg, O., & Van Rijn, H. (2023). Improving Adaptive Learning Models Using Prosodic Speech Features. In: Wang, N., Rebolledo-Mendez, G., Matsuda, N., Santos, O.C., Dimitrova, V. (eds) Artificial Intelligence in Education. AIED 2023. *Lecture notes in computer science, 13916* (pp. 255–266). https://doi.org/10.1007/978-3-031-36272-9_21

Wilschut, T., Sense, F., & van Rijn, H. (2025). Modality Matters: Evidence for the Benefits of Speech-Based Adaptive Retrieval Practice in Learners with Dyslexia. *Topics in cognitive science*, *17*(1), 57–72. https://doi.org/10.1111/tops.12769

Wilschut, T., Sense, F. & van Rijn, H. (In review). Cognitive and Metacognitive Markers of Memory Retrieval Performance in Speech Prosody.

Appendix

Subject-Verb Sentences Used as Experiment Stimuli

Table A1

40 Short Subject-Verb Sentences Created for the Language Learning Task

ID	Dutch Sentence	Number of Syllables	English Translation	Italian Sentence	Number of Syllables
1	De nicht glimlacht	4	The niece smiles	La nipote sorride	7
2	De baby eet	4	The baby eats	Il bambino mangia	6
3	De kikker springt	4	The frog jumps	La rana salta	5
4	Het meisje leest	4	The girl reads	La ragazza legge*	6
5	De deur kraakt	3	The door creaks	La porta scricchiola	6
6	De broer rijdt	3	The brother drives	Il fratello guida*	6
7	De trein wacht	3	The train waits	Il treno aspetta*	6
8	De zus loopt	3	The sister walks	La sorella cammina*	7
9	De hond speelt	3	The dog plays	Il cane gioca*	5
10	De schilder tekent	5	The painter draws	Il pittore disegna*	7
11	De nacht begint	4	The night starts	La notte inizia*	6
12	De brandweerman spreekt	5	The firefighter speaks	Il pompiere parla*	6
13	De man liegt	3	The man lies	L'uomo mente*	4
14	De prijs verandert	5	The price changes	Il prezzo cambia	5
15	De vrouw komt	3	The woman comes	La donna viene*	5
16	Het bot breekt	3	The bone breaks	L'osso si rompe	5
17	De klok tikt	3	The clock ticks	L'orologio ticchetta	7
18	De boom bloeit	3	The tree blooms	L'albero fiorisce	6
19	De dochter schreeuwt	4	The daughter shouts	La figlia grida	5
20	De vriend lacht	3	The friend laughs	L'amico ride	5
21	De vogel zingt	4	The bird sings	L'uccello canta	5
22	De groep wint	3	The group wins	Il gruppo vince	5

ID	Dutch Sentence	Number of Syllables	English Translation	Italian Sentence	Number of Syllables
23	De haai zwemt	3	The shark swims	Lo squalo nuota	5
24	De zon schijnt	3	The sun shines	Il sole splende	5
25	De familie betaalt	6	The family pays	La famiglia paga	6
26	De leeuw slaapt	3	The lion sleeps	Il leone dorme	6
27	De maan bestaat	4	The moon exists	La luna esiste	6
28	De stoel valt	3	The chair falls	La sedia cade	5
29	De muis verdwijnt	4	The mouse disappears	Il topo scompare	6
30	De kat snurkt	3	The cat snores	Il gatto russa	5
31	De buurman begrijpt	5	The neighbour understands	Il vicino capisce	7
32	Het hout brandt	3	The wood burns	Il legno brucia	5
33	De plant groeit	3	The plant grows	La pianta cresce	5
34	De docent vermenigvuldigd	8	The teacher multiplies	L'insegnante moltiplica	8
35	De jongen rent	4	The boy runs	Il ragazzo corre	6
36	Het dier reist	3	The animal travels	L’animale viaggia	6
37	De ober kookt	4	The waiter cooks	Il cameriere cucina	8
38	De oma bakt	4	The grandma bakes	La nonna inforna	6
39	Het vat explodeert	5	The barrel explodes	Il barile esplode	7
40	Het schip zinkt	3	The ship sinks	La nave affonda	6

Note. The two groups of sentences used for the experimental task (and counterbalanced between conditions) contain items 1 to 15 and 16 to 30, respectively. Sentences 31 to 34 were presented in practice trials, and the remaining ones were not put to use. Asterisks in some Italian translations indicate that the spoken stimuli for these sentences had a different voice than the rest of the spoken stimuli.