Supplementary Materials

Automatic Podcast Transcription

To help characterize the statistics of the test words within the podcasts and support subsequent analyses (described further below), automatic transcriptions of the Italian podcasts were generated using the transcription function of *YouTube*. As a verification step, we quantified the accuracy of these automatic transcription by comparing them to a human transcription of the podcast samples, performed by a native Italian speaker. Our native Italian speaker manually transcribed a 20-minute sample of both *Parole di Storie and Radio Feltrinelli* (40 min total), which we then compared directly to the automatic transcriptions. Words that were absent or written incorrectly in the automatic transcriptions were coded as 0; words that were correctly transcribed were coded as 1. Calculated in this way, word-by-word accuracy of the automatic transcription was 96.2% for *Parole di Storie* and 90.7% for *Radio Feltrinelli*. While accuracy was thus reasonably good in both cases, we note that it may have been somewhat lower for *Radio Feltrinelli* as this podcast contains free-flowing and spontaneous speech, rapid turns in conversation, interruptions and periods where two talkers are speaking at once, stuttering, etc., which can be challenging for even a professional transcriber to represent. In contrast, *Parole di Storie* consists of rehearsed speech produced by a single narrator. Thus, although not 100% accurate, the automatic transcriptions provide a reasonable estimate of L2 participants' listening experience, particularly for speech that is produced in the clear.

Characterization of Test Words

Transitional Probabilities

The words in each category were selected to ensure that the transitional probabilities (TPs) of neighbouring syllables within the true Italian words (both trained and nontrained words) were higher than those within the foil items. As described previously, TP is the probability that one syllable will occur, given the presence of the immediately preceding syllable (e.g., in "baby", the probability that "by" follows "ba," given that "ba" has occurred; see Saffran, Newport, et al., 1996). We estimated the TPs of syllables that were characteristic of L2 participants' listening experience by using the automatic transcriptions of the 14 Italian podcasts. We used the basic orthographic representations for our estimates because Italian is considered to have a shallow orthography, such that a given letter of the alphabet is almost always pronounced the same way, regardless of the word it appears in. Thus, in general, computing the TPs of a given word based on its orthography provides a good estimate of the TPs of its spoken syllables. Because the transcripts did not afford access to information about the presence of pauses, TPs were estimated solely on the basis of word-internal TPs of words present in the podcast (i.e., computed under the assumption that each individual word was spoken in isolation and not connected to the preceding and subsequent words). However, a subsequent analysis (see below) found that whether or not pauses were included in this computation appeared to have little impact on test word TP estimates.

First, the 14 automatic transcriptions were concatenated together, providing a list of all individual words that were presented during the podcasts. Next, for a given candidate test word (e.g., *bas.sot.to*), the TP for each syllable pair (e.g., *bas.sot*) was calculated by dividing the frequency of the syllable pair in the concatenated transcripts by the frequency of the initial syllable alone (e.g., frequency of *bas.sot* / frequency of *bas*). Syllabification of test words was carried out manually on a word-by-word basis, based on established syllabification rules in

Italian (e.g., see www.italianlanguageguide.com/pronunciation/syllabification.asp). As a measure of a word's overall transitional probability, a "total transitional probability" (TTP) value was then calculated by adding together the transitional probability of the first (TP1) and second (TP2) syllable pair (e.g. word = a.ba.ca, TP1 = (a.ba/a), TP2 = (ba.ca/ba), TTP = TP1 + TP2). An item analysis confirmed that the TTPs between trained and nontrained words did not differ significantly (trained word mean = 0.16; SD = 0.29; nontrained word mean = 0.10; SD = 0.15; $F(1,38) = 0.66$; $p = 0.41$), while the TTPs for foil items were significantly lower than the TTPs for the true Italian words (mean TTP foil words = 0.011; SD = 0.037; $F(1,58) = 5.34$, $p = 0.024$). Within each category (trained words, nontrained words, foil words), TTPs did not differ significantly between the two sentences sets (A versus B; all p values > 0.26).

As previously described, TPs were estimated solely on the basis of word-internal transitional probabilities of words present in the podcast. To examine the potential impact of pauses on TP estimates, we computed TPs for each of the 60 test items under the assumption that all words within each podcast were connected (i.e., we removed all pauses between individual words in the transcriptions). TP estimates for test words were extremely highly correlated with the original TP estimates ($r = 0.97$, $p < 0.001$), suggesting that taking into account presence or absence of pauses between words did not have a large effect on TP estimates. Thus, we conclude that a more rigorous estimate of pauses in the transcript would likely would have had only a minimal impact on the TP estimates.

Individual Syllable Occurrences

As a point of comparison, we also calculated the total number of individual syllable occurrences within the podcasts for each of the 60 test words. The number of occurrences for the three syllables were then summed together to produce a total syllable occurrence score for each

test word (e.g., total individual syllable occurrences for *bas.sot.to* = number of occurrences of *bas* + number of occurrences of *sot* + number of occurrences of *to* within the podcasts).

Test Word Occurrences

Finally, for trained and nontrained test words, we quantified the number of presentations of each word throughout the 14 Italian podcasts. This information was used in subsequent analyses in order to address whether any increase in familiarity for true Italian words was driven by the learning of specific words that had appeared more frequently in the podcasts. Again, the number of word occurrences of each of the 40 true Italian words was tallied from the automatic podcasts transcriptions. Most of the test words either never appeared in the podcasts (32/40 words) or appeared very rarely (1-3 times; 5/40), while three words (*bambino, piacere,* and *ragazzo*) appeared more frequently (11-32 times). Number of test word exposures did not significantly differ between word sets (A/B; $F(1,36) = 0.41$, $p = 0.53$) or between trained versus nontrained words ($F(1,36) = 1.40$, $p = 0.24$).

Baseline Familiarity

We further confirmed that the trained and nontrained test words were well-matched on overall baseline familiarity by collecting Italian word-like ratings from a separate, control group of participants ($n = 124$). As in the main experiment, participants in this group were all self-identified native English speakers with no significant Italian experience. Participants were asked to indicate to provide ratings on a 1-4 scale for each of the 60 items used in the task, indicating how much like a real Italian word each item sounded to them (4 being the most likely to be an Italian word). Importantly, ratings to trained and nontrained stimuli were virtually identical (trained mean rating = 2.86, SD = 0.33; nontrained mean rating = 2.86, SD = 0.35; Word

Category Effect: $F(1,22) = 0.004$, $p = 0.95$). In contrast, foil items were rated as significantly less Italian-like than the trained and nontrained words (mean rating = 2.63, SD = 0.37, Word Category Effect: $F(2,44) = 51.6$, $p < 0.001$). The lower mean baseline rating for foils is likely a result of having selected these items to contain lower internal transitional probabilities in Italian, such that they may sound inherently less word-like or Italian-like. This is not inherently an issue, as our design allows us to disentangle effects of baseline familiarity and L2 exposure by measuring familiarity ratings both prior to and after the 2-week listening period.

Podcast Questionnaire

Participants were required to report on the secret words embedded in each podcast on each day, allowing us to track participant compliance with the daily listening protocol.

Analysis

For each of the 14 daily podcast questionnaires, participants' responses to each item were coded as 1 for correct (indicating that a "secret" word that had appeared in the podcast was correctly endorsed, or a that word that had not appeared in the podcast was correctly rejected) and 0 for incorrect. For each participant, the overall response accuracy was then calculated by dividing the number of correct answers by the total number of secret words embedded within all 14 podcasts. Performance on the questionnaire was used to confirm that individuals had complied with the 14-day listening protocol. An independent samples t-test was used to test whether podcast accuracy differed significantly between the two groups.

Results

Podcast questionnaire responses confirmed participants' general compliance with the 14-day listening protocol. The average rate of correct responses was 95.8% (SD = 8.7%) for participants in the Italian-exposure group and 90.0% (SD = 10.4%) for participants in the control group. Participants in the L2 exposure group significantly outperformed those in the control group ($t(61) = 2.72$, $p = 0.009$, $d = 0.083$).

Daily Listening Habits (Exit Questionnaire)

Participants reported doing a number of different daily activities while listening to the podcasts. Common responses included going for a walk; exercising; completing household tasks such as cleaning, laundry, and preparing meals; eating; surfing the Internet or scrolling through social media; playing video games; relaxing/hanging out at home; studying; reading; and driving or commuting on the bus.

Exposure + Word Detection Task Results

Word Detection Rate

Across both groups, the mean word detection rate was 87.8% (SE = 1.5%) in Session 1 and 84.6% (SD = 1.5%) in Session 2. This small decline in performance from Session 1 to Session 2 was significant, though unexpected ($F(1,6246) = 14.3$, $p < 0.001$). There was no overall effect of Group on word detection rate ($F(1,6246) = 0.021$, $p = 0.88$), nor was there any significant interaction between Group and Session ($F(1,6246) = 1.55$, $p = 0.21$).

Response Time

Across both groups, the estimated marginal mean response time to identified words was 1034 ms (SE = 19.2 ms) in Session 1 and 1051 ms (SE = 19.4 ms) in Session 2. The effects of Session,

Group, and interaction between Session and Group on response times were all not significant (all *p* values > 0.26). These results indicate that participants were capable of detecting specific words embedded in continuous, fluent Italian speech, regardless of experience with Italian speech. Listening to Italian podcasts did not significantly boost this ability.

False Alarms

False alarms were rare, occurring an average of 3.7 times (SE = 0.56) per session (i.e., across 300 total sentences). The large majority of these false alarms occurred within training sentences (mean = 3.2 times per session). False alarms occurring at an incorrect position within a target sentence occurred very infrequently (mean = 0.53 times per session).

Supplementary Discussion of Exposure + Word Detection Task Results

Trained L2 words are accurately detected and recognized, independently of L2 listening experience

A secondary result was that participants were generally successful in detecting and recognizing trained words, regardless of L2 listening experience. On the exposure + word detection task, word detection performance was quite good across sessions and groups (>85%). Similarly, on the familiarity rating task, participants successfully recognized previously encountered words, rating them as significantly more familiar than matched nontrained words. L2 exposure did not improve either online word detection or rating performance for trained words, with no significant interactions between group and session.

The effect of recent exposure on word identification conceptually replicates a previous study (Kittleson et al., 2010). After being briefly familiarized with Norwegian speech, non-

speakers of Norwegian were able to differentiate between words embedded in the speech stream and nonwords. Interestingly, this ability was not modulated by participants' language background. Taken together, these results suggest that the ability to segment and identify individual words in continuous speech may be supported at least partially by domain-general perceptual and memory mechanisms, which are not dependent on expertise with a particular language. However, while L2 exposure did not improve word detection or trained word endorsements in the current study, it remains possible that some improvement on these measures may be observed with additional L2 exposure (e.g., months of training rather than weeks).