Semantics-weighted lexical surprisal modeling of naturalistic functional MRI time-series during spoken narrative listening

Andrea G. Russo\textsuperscript{a,b,}\textsuperscript{*}, Maria De Martino\textsuperscript{a}, Azzurra Mancuso\textsuperscript{a}, Giorgio Iaconetta\textsuperscript{b,}\textsuperscript{c}, Renzo Manara\textsuperscript{b,}\textsuperscript{c}, Annibale Elia\textsuperscript{a}, Alessandro Laudanna\textsuperscript{a}, Francesco Di Salle\textsuperscript{b,}\textsuperscript{c}, Fabrizio Esposito\textsuperscript{b,}\textsuperscript{c}

\textsuperscript{a} Department of Political and Communication Sciences, University of Salerno, Fisciano, Salerno, Italy
\textsuperscript{b} Department of Medicine, Surgery and Dentistry, "Scuola Medica Salernitana", University of Salerno, Baronissi, Salerno, Italy
\textsuperscript{c} Department of Diagnostic Imaging, University Hospital "San Giovanni di Dio e Ruggi D'Alagona", Salerno, Italy

\textbf{A R T I C L E   I N F O}

\textbf{Keywords:}
Naturalistic stimulus
Surprisal
Narrative comprehension
Functional MRI
Language
Italian

\textbf{A B S T R A C T}

Probabilistic language models are increasingly used to provide neural representations of linguistic features under naturalistic settings. Word surprisal models can be applied to continuous fMRI recordings during task-free listening of narratives, to detect regions linked to language prediction and comprehension. Here, to this purpose, a novel semantics-weighted lexical surprisal is applied to naturalistic fMRI data.

fMRI was performed at 3 Tesla in 31 subjects during task-free listening to a 12-minute audiobook played in both original and word-reversed (control) version. Lexical-only and semantics-weighted lexical surprisal models were estimated for the original and control word series. The two series were alternatively chosen to build the predictor of interest in the first-level general linear model and were compared in the second-level (group) analysis.

The addition of the surprisal predictor to the stimulus-related predictors significantly improved the fitting of the neural signal. In average, the semantics-weighted model yielded lower surprisal values and, in some areas, better fitting of the fMRI data compared to the lexical-only model. The two models produced both overlapping and distinct activations: while lexical-only surprisal activated secondary auditory areas in the superior temporal gyri and the cerebellum, semantics-weighted surprisal additionally activated the left inferior frontal gyrus.

These results confirm the usefulness of surprisal models in the naturalistic fMRI analysis of linguistic processes and suggest that the use of semantic information may increase the sensitivity of a probabilistic language model in higher-order language-related areas, with possible implications for future naturalistic fMRI studies of language under normal and (clinically or pharmacologically) modified conditions.

\textsuperscript{*} Corresponding author at: Department of Political and Communication Sciences, University of Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano, Italy.
\textit{E-mail address: andrusso@unisa.it} (A.G. Russo).

https://doi.org/10.1016/j.neuroimage.2020.117281
Received 22 March 2020; Received in revised form 22 June 2020; Accepted 13 August 2020
Available online 21 August 2020

1053-8119/© 2020 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license. (http://creativecommons.org/licenses/by/4.0/)
ness (i.e. the surprise) of this word, given the current context (as determined by current and previous words), and it has been parametrically linked to the language-related cognitive effort or the linguistic processing difficulty (Demberg and Keller, 2008; Hale, 2016, 2001; Levy, 2008; Smith and Levy, 2013), a possible correlation between the participants’ elicited BOLD response (taken as a neural index of cognitive effort) and the participants’ scores to the post hoc questionnaire (taken as a behavioral index of their comprehension of the story) was also hypothesized.

2. Materials and methods

2.1. Participants

Thirty-one healthy volunteers (Italian native proficient speakers, 23 females, mean age 24.2 ± 4.4 years old) without known psychiatric or neurological problems, with normal or corrected-to-normal vision and without hearing, developmental and language-related problems, were enrolled in the experiment. All participants were right-handed by self-report, and all participants were naive with respect to the purpose of the experiment. Written informed consent was obtained in accordance with the Declaration of Helsinki, and the study was approved by the local ethics committee.

2.2. Stimuli

The stimulus used in the present study has been taken from the website of “Progetto Babele Rivista Letteraria” (http://www.progettobabele.it) where several written and spoken Italian narratives of semi-professional and amateur writers are freely available. The excerpt used in this work is “Storia di Gianna e delle sue chiavi” (“Story of Gianna and her keys”) written by Carlo Santulli and read by Silvia Cecchini (https://www.progettobabele.it/AUDIOFILES/ascolta.php?ID=841).

The excerpt was spoken at a normal rate: an average rate of 156 words per minute that fits in the recommended word per minute rate for audiobooks (Williams, 1998). Stimulus duration was 11:50 min (1878 written words).

The bigrams with the presence of an apostrophe were entered in the model first as two separate words, then only the surprisal relative to the second word was considered as it includes the whole bigram and it was listened by the subjects as a single word. Therefore, the final number of words used in the subsequent analysis was 1856.

Reversed speech version of the stimulus was obtained with Audacity 2.03 (https://www.audacityteam.org/). Briefly, the original audio track was imported in Audacity and then a transformation was applied so that the end of the audio was heard first and the beginning last. This technique has been used in previous studies (Lopopolo et al., 2017; Willems et al., 2016) as the reversed speech is comparable to forward speech in terms of auditory characteristics while omitting the linguistic components (but see also Brown et al., 2012). To avoid and reduce possible biases due to previous knowledge of the story, in this study a short narrative of an amateur writer was chosen as our stimulus. Indeed, all subjects declared that they did not have any prior knowledge of the story.

2.3. Definition of surprisal

A human speaker, using specific rules (syntax) sets the order of a list of words to convey a message. As a consequence, it can be postulated that the language system, after the processing of the first t−1 words (W1, . . . , Wt−1) of a stream will be in a state that implicitly assigns a conditional probability: P(Wt|W1, . . . , Wt−1) to each potentially occurring word Wt (Kuperberg and Jaeger, 2016).

The surprisal associated with an upcoming new word appearing at time (or position) t is defined as the negative logarithm of its conditional occurrence probability:

\[
\text{surprise}(t) = - \log_{10}(P(W_t|W_1, \ldots, W_{t-1}))
\]

(1)

If the conditional probability of the observed word is one it means that given the left-side context of the word there are no other possible outcomes, thereby the surprise in observing the word is null. On the
other hand, the occurrence of a word that was not among the candidate outcomes, i.e. an event with a null probability, corresponds to infinite surprisal value.

The conditional probabilities needed to estimate the surprisal values can be calculated by any probabilistic language model (Henderson et al., 2016; Willems et al., 2016).

2.4. Definition of semantic vector space

In 1954, Harris postulated the distributional hypothesis of language affirming that the meaning of a word can be inferred from the contexts in which it is used (Harris, 1954). Using word co-occurrences in a large corpus, it is possible to observe that, for example, the contexts in which the word “merchant” is used are similar to those in which the word “dealer” occurs, whereas the context of occurrence of “dog” and “pillars” are essentially different (Triponi and Pira, 2017). In the natural language processing field, this hypothesis has driven the creation of several algorithms to create vector space models where each word is represented by an N-dimensional vector (Dumais, 2004; Mikolov et al., 2013; Pennington et al., 2014; Sayeed et al., 2015; Turney and Pantel, 2010), where N is the number of the so-called context words that are the most common words in the language corpus (excluding words such as the determiners and conjunctions).

The $i$-th component of the word vector $w$ (ci) is estimated as the ratio between the conditional probability of the context word $c_i$ given the word $w$ and the (unconditional) probability of the context word $c_i$

\[
 v_i = \frac{p(c_i|w)}{p(c_i)} \tag{2}
\]

The numerator in (2) can be obtained by iterating through the corpus and counting how many times word $w$ appears together with a context word $c_i$ within a fixed window of words. A window size of 10 was selected to weaken possible syntactic effects and the 1000 most common words (excluding function words such as determiners) of the corpus were chosen as context words (Huth et al., 2016a; Sayeed et al., 2015). Vector representations of unknown words (i.e. words not present in the corpus) were artificially created using a vector of ones to guarantee the mathematical structure of the model and to not bias these unknown words towards a specific context word.

This representation allows estimating the semantic similarity between two words using the cosine of the angle between the vector words $\hat{v}$ and $\hat{h}$

\[
 \cos(\phi) = \frac{(\hat{v} \cdot \hat{h})}{|\hat{v}||\hat{h}|} \tag{3}
\]

In this work to prepare the corpus and to estimate the vector semantic space a combination of custom scripts in Python (Software Foundation. Python Language Reference, version 3.5. available at http://www.python.org) was used based on the package Natural Language Toolkit (NLTK) (Bird et al., 2009) and TreeTagger (Schmid, 1994). Additional custom scripts in R (Tierney, 2012) and bash scripts were also used.

2.5. Estimation of the surprisal models

Two different types of surprisal were estimated. The first is based on the word co-occurrences of the observed lexical form given the preceding local context estimated on a large corpus, therefore it can be referred to as a lexical surprisal (LS). In the second type, a multiplicative factor expressing the semantic similarity of the current word with a preceding broader context (independent from the local one) is introduced to modulate the co-occurrences of the current lexical form with the left-side local context, using the distributed representations of the words in a vector space model, therefore it can be called semantics-weighted surprisal (SwS).

The word co-occurrences used to build both the semantic vector space model and the probability for the LS need to be estimated on a large linguistic corpus. In this study, the PAISÂ corpus that is a Creative Commons licensed large web corpus of contemporary Italian was used. The latter corpus contains approximately 388,000 documents from 1,067 different websites, for a total of about 250M tokens (Lyding et al., 2014).

Word co-occurrences were estimated after a series of pre-processing steps on the corpus that was necessary in order to fulfill the requirements of the used software (see next sections), i.e. to reduce the sparsity of the word vector and the computational cost. Thereby, residuals and spurious HTML tags, words with a total frequency of less than fifty (in most cases they represented typos or obsolete words) were removed from the corpus. Finally, the corpus was split into sentences (one sentence per line), all words were converted to lower case, and the punctuation marks were removed.

2.5.1. Estimation of the lexical surprisal

The LS was based only on the co-occurrences of the observed lexical form given its context in a large corpus. In this work, a stable and widely used stochastic model was used, the second-order Markov model, also known as trigram model (Armeni et al., 2017; Willems et al., 2016). It is based on the idea that, in a given sentence formed by words $W_1, W_2, \ldots, W_T$, the probability of observing the word $W_i$ given the whole left-side context $P(W_i|W_{i-2}, W_{i-1})$ can be approximated to $P(W_i|W_{i-2}, W_{i-1})$ (i.e. the probability of observing the word $W_i$ given only the two preceding words). Surprisal values estimated by trigram models have been used successfully in recent neuroimaging studies (Frank and Willems, 2017; Loppopolo et al., 2017; Willems et al., 2016) and in many psycholinguistic EEG studies that demonstrated that trigram-based surprisal correlates positively with the N400 (Frank et al., 2015). The LS values were estimated using the software SRILM and Kneser–Ney smoothing was used to control for possible unknown words, i.e. words that are not present in the corpus (Stolcke et al., 2011). Unknown words were either atypical inflected forms of an Italian word or obsolete words. The absence in the corpus was either due to the corpus preprocessing (e.g. the removal of words with a frequency lower than 50) or to the fact that these words were already absent in the web documents that were searched during the creation of the corpus (Lyding et al., 2014). In total, after corpus preprocessing, 3.7% of the words in the corpus were labeled as unknown. For example, an unknown (obsoleto) word was the word “vapoforno” (literally steam-oven) that refers to a specific oven used in the past to bake the bread, whereas a word with a frequency of 1, and thus lower than 50, was “sbattei” (I slammed).

2.5.2. Estimation of the semantics-weighted surprisal

The SwS was estimated following the semantic surprisal model originally presented in Mitchell and Lapata (2009) and recently implemented and used by Sayeed and colleagues in a psycholinguistic study (Sayeed et al., 2015). In this study, it has been shown that the SwS can successfully predict spoken word durations in naturalistic scenarios such as groupwork meetings. Briefly, this model integrates the probabilities obtained by the trigram model and the semantic similarity calculated from the vector space model between the considered word and its preceding words (i.e. the history of the word). However, a distinction between the “content” words (i.e. words that carry relevant semantic meaning such as “cat” or “dog”) and “function” words (i.e. words that do not have a clear semantic meaning such as the determiners) is made. For content words, the trigram probability is scaled by a positive factor depending on the semantic similarity of the current word with its recent history (i.e. the words constituting the preceding context) (Mitchell and Lapata, 2009). The theoretical range of this factor is between zero and infinite and therefore it can either result in a downscaling (if it is lower than 1) or an upscaling (if it is higher than 1) of the trigram probability. For function words, the semantic scaling factor is simply set to 1. Therefore, assuming that for the function word the SwS is $P(W_i|W_{i-2}, W_{i-1})$ (i.e. it equals LS), for the content words the SwS is estimated as
in the MRI console via loudspeakers connected to the stimulation computer. Before the beginning of the acquisition, a volume test was performed: a fragment from another story with comparable voice and sound quality was presented while the scanner was collecting images. The volume of the audio was adjusted to the optimal level based on feedback from the participant.

2.7. **Post-hoc questionnaire**

After the acquisition, participants underwent a questionnaire to test their comprehension of the non-reversed story. The questionnaire contained 8 multiple choice questions regarding the non-reversed story with 4 answer options to each question. Questions were about general content, and correct answers were summed to have an overall level of understanding and attention of each participant.

2.8. **Image acquisition**

MRI was performed on a 3T scanner (Magnetom Skyra, Siemens Healthcare, Germany) equipped with a 20-channel parallel head coil. The fMRI scan consisted of 750 volumes of a multi-band (Feinberg et al., 2016; Moeller et al., 2010; Xu et al., 2013) repeated gradient-echo echo-planar imaging (EPI) sequence (repetition time (TR) = 1000 ms, echo time (TE) = 30 ms, number of axial slices = 60, matrix = 96 × 96, field of view (FOV) = 240 mm, thickness= 2.5 mm, interslice gap = 0 mm, multi-band factor = 4). Two more multi-band gradient-echo EPI sequences (5 volumes each) were acquired. The first was identical to the sequence of 750 volumes, whereas the second was acquired with an opposite phase encoding. Three-dimensional T1w Magnetization Prepared Rapid Gradient-Echo (MPRAGE) sequence (TR =2400 ms, TE= 2.26 ms, TI = 950 ms, flip angle =8°, slice thickness = 1.0 mm, matrix size = 256 × 256, number of slices = 192 and voxel size = 1.0 × 1.0 × 1.0 mm³) was acquired in the same session in order to have a high resolution anatomical reference for registration and normalization of the functional images.

2.9. **Prediction performances**

When comparing the prediction performances between a pure lexical and a semantics-weighted surprisal model, it is important to consider that (i) the Sws model integrates information of different nature (i.e. word co-occurrences and semantic vectors) and (ii) the words included in the conditioning context are not the same in the two models.

Regarding the first point, it is worth noting that the Sws is a weighted version of the LS where the semantic of the word is explicitly modeled, whereas, in the LS model, the semantic of the word is still present in the lexical form but it is not explicitly taken into account during the surprisal estimation.

Regarding the second point, Frank and Willems (2017) introduced a modified lexical surprisal model, called “skip-bigram”, to overcome this disparity in the conditioning context. In brief, the “skip-bigram” probability ($P_{SB}$) is estimated from the co-occurrence of the pair (bigram) formed by the considered word and each word outside the context used for the trigram model, that has been used to estimate the semantic distance. The estimated probabilities are then interpolated with the lexical probability (see Frank and Willems, 2017 for more details). According to this study, an adapted version of this model could reduce the influence of the length of the conditioning domains (see Section 2.5.2) in the comparison as it takes into account the same context of the semantic distance that is based (by definition) on a higher number of context words (four context words vs. two context words). Therefore, as the difference between the LS and the Sws is in how they model the word meanings, the LS interpolated with the $P_{SB}$ allows us to investigate and to isolate the “semantics-weighting” as it made the two conditioning domains between the pure lexical and the semantics-weighted model perfectly equitable.

2.6. **Experimental procedure**

Participants listened to a story, as well as to its reversed version, while in the MRI scanner. Half of the participants started with the non-reversed stimulus and half with the reversed speech stimulus. Before entering the scanner, the participant was instructed to listen as carefully as possible. A short break separated the two versions of the stimulus.

Stimuli were presented by means of custom scripts written in Python 2.7 with the use of the PsychoPy2 module (Peirce, 2008, 2007) via MRI compatible earphones (Serene Sound, Resonance Technology, USA). For online monitoring of the experiments, the stimuli were also reproduced in the MRI console via loudspeakers connected to the stimulation computer. Before the beginning of the acquisition, a volume test was performed: a fragment from another story with comparable voice and sound quality was presented while the scanner was collecting images. The volume of the audio was adjusted to the optimal level based on feedback from the participant.
The “skip-bigram” factor for each word was estimated by considering the same content words used for the estimation of the SwS semantic component. Thus, a modified version of the LS, called “skip-bigram” LS (LS\textsubscript{SG}), was estimated by linearly interpolating the \( P_{SB} \) with the LS:

\[
LS_{SB} = - \log \left( \lambda P_{W} \log W_{i-1}, W_{i-2} \right) + (1 - \lambda) P_{SB}
\]

where \( \lambda \in [0, 1] \) indicates the contribution of the \( P_{SB} \) to the surprisal. The lower the value in this modified surprisal, the better is the prediction (Frank and Willems, 2017). Therefore, the \( \lambda \) factor that minimized the \( LS_{SB} \) average was chosen.

An estimate of the average surprisal was computed for (and compared between) the three models (SwS, LS and \( LS_{SB} \)), by averaging the surprisal values across all words in the narrative text.

**2.10. Functional MRI data analysis**

MRI data pre-processing was performed using BrainVoyager (Brain Innovation, The Netherlands, www.brainvoyager.com) (Goebel, 2012), SPM12 (Wellcome Department of Imaging Neuroscience, London, UK, http://www.fil.ion.ucl.ac.uk/spm/), MATLAB R2018b (The MathWorks, Inc., Natick, MA, www.mathworks.com) and FSL (Jenkinson et al., 2012). FMRI statistical analysis was performed in BrainVoyager.

**2.10.1. FMRI data preprocessing**

The raw slice time-series of each scan (DICOM series) were first imported in BrainVoyager and carefully reviewed with the time-course movie tool to promptly detect the occurrence of spikes in the images or other acquisition-related technical problems. In BrainVoyager, slice timing correction, followed by motion correction and high-pass filtering (cut-off to 0.008 Hz) was applied to all fMRI time-series. Moreover, starting from the motion parameters, the maximum absolute translation/rotation values and the mean framewise displacement (FD) were estimated. No de-spiking was applied to the fMRI time-series, which were exported to NIfTI format and corrected for EPI geometrical distortions with the FSL tool TOPUP (Andersson et al., 2003; Jenkinson et al., 2012; Smith et al., 2004) using the two additional short fMRI time-series acquired with opposite phase encoding to estimate the warps. The corrected series were then further pre-processed using the Data Processing Assistant for Resting-State FMRI toolkit (DPARSF; http://www.fmri.org) (Chao-Gan and Yu-Feng, 2010). In particular, the anatomical and functional scans were spatially normalized to the standard Montreal Neurological Institute (MINI) template, functional images were resampled to 2 × 2 × 2 mm voxel sizes and then, white matter (WM), cerebrospinal fluid (CSF) signals and the Friston-24 movement parameters (Friston et al., 1996) were regressed out from fMRI time series. Finally, functional images were spatially smoothed using an isotropic 4-mm full-width at half-maximum Gaussian kernel.

After these steps, all individual fMRI series were imported back in BrainVoyager and further transformed into the Talairach space. This step allowed the statistical analysis and presentation of group-level activation maps in the surface space on the cortical surface meshes.

**2.10.2. FMRI data modeling and statistical analysis**

For the whole-brain voxel-based statistical analysis of the fMRI time-courses, a two-level (mixed-effects) general linear model (GLM) was performed in BrainVoyager. In the first-level GLM, the correlation between surprisal estimates and fMRI time-courses was estimated as a fixed effect in every single subject in both narrative conditions (original and reversed speech). In the second-level GLM, the inter-subject variability of these effects was assessed by treating subjects as random observations.

The first-level general linear model (GLM) was applied to the volume time-courses of each subject (Friston et al., 1995) alternatively accounting for the LS or the SwS of the words. In addition to the predictor of interest, three predictors of no interest (confounds) accounting for word duration (WD), lexical frequency (LF) and root mean squared (RMS) amplitude of the word sound (WS) were added to the design matrix. The inclusion of the LF ensures that no extra effects result from the correlation of fMRI responses with LS and SwS predictors because of the absolute (context-independent) rarity of a word in a language-representative corpus, whereas the inclusion of the WD and the RMS of the WS in two additional parametric nuisance predictors ensures that no extra effects result from purely acoustic sources of variance, such as the (changing) volume of the sound, or the variable prosody, intonation and tone of the speaker’s voice. To justify the use of separate fMRI data models for the two surprisal predictors (LS, SwS), as well as the inclusion of the LF as confound predictor in both models, some additional analyses were performed to explore in greater details the multicollinearity of all three predictors (LS, SwS, LF) as well as the neural effects associated with the LF variable. The description and the results of these analyses are provided in Section 1 of the supplementary material.

For each single word, starting from the original speech, LS and SwS values were estimated using a custom Python script. The LS and SwS series were reversed for the reversed speech. Word onsets (times-tamps) and durations were separately calculated for the original and reversed speech using Speechmatics (https://www.speechmatics.com/). The (log-transformed) LF of each word was obtained from the PAIS\textalpha corpus (Lyding et al., 2014). As the single word represents the linguistic unit of interest for our model, all parameters (SwS, LS, WD, LF and RMS) were related to the single words, thereby a parametric design was used to build the predictors in the GLM. To this purpose, word onsets and durations were used to create a box-car predictor, i.e. a function that is one for the whole duration of a word and zero otherwise, at the time resolution of 0.01 s (which is the time resolution of the word times-tamp). To generate the LS, SwS and LF predictors, the boxcar predictor was amplitude-modulated at each word by the value of the parameter at that word. To generate the WS predictor, the RMS value of the audio signal was estimated over the time interval between the onset and the offset of each word. Before modulation, the series of each parameter were converted to z scores. The unmodulated (box-car) predictor was also taken to account for the variable durations of each word. All four predictors were then convolved with the hemodynamic response (to account for the hemodynamic delay) at the sampling rate of 100Hz.

Finally, the convolved predictors were down-sampled at the time resolution of fMRI (1 Hz). This procedure was applied identically for the original and reversed speech stimuli. As the two surprisal models were applied separately, two separate GLM were performed where the predictor of interest was either the LS or the SwS predictor. To decorrelate the other (confound) predictors, the Gram-Schmidt procedure was implemented according to its hierarchical formulation, whereby the first predictor is simply unaffected by the procedure, the second predictor is orthogonal to the first, the third predictor is orthogonal to the subspace spanned by the first two predictors, and so on. This is equivalent to replacing each confound variable by its residuals from a least squares regression on the previous variables. Because each subject performed two runs (in random order), the modeled time courses from both runs (original and reversed speech) were concatenated along time and the two design matrices were both concatenated along the time dimension (rows) and duplicated (with zero-padding) along the predictor dimension (columns) to keep the original and reverse speech conditions separate.

For the second (group) level analysis, two separate random-effects (RFX) GLM analysis were performed: one with the LS and one with the SwS, as predictor of interest in the first-level GLM. In both cases, the contrast between the “original” and the “reversed” narrative stimulus condition was evaluated, resulting in two statistical maps (t statistics). These maps were overlaid in pseudo-color on, and for visualization purposes, projected on an inflated cortical mesh of a Talairach-normalized T1-weighted scan of a single subject. To correct for the multiple voxel-wise comparisons, a cluster-level thresholding approach was applied (Forman et al., 1995; Goebel et al., 2006), thereby the t maps were
initially thresholded at a maximum voxel-wise \( p \)-value of 0.001 (uncorrected) and then given in input to a whole-brain (no mask) correction procedure based on the estimate of the intrinsic spatial smoothness of the map and on 1000 iterations of a Monte Carlo simulation, to determine the minimum cluster size threshold ensuring a corrected \( p \)-value of 0.05 (cluster-level corrected) at each voxel.

2.10.3. Likelihood-based model comparison and model fitting

The clusters where the surprisal models elicited significant activations in the original speech compared to the reversed speech were further investigated to understand how well the two surprisal models explain the variance of the BOLD signal. In particular, the average BOLD time-course for both the original and the reversed story were extracted from these clusters in each subject’s data set and the resulting inter-subject data matrix (N subjects by 750 × 2 volumes) was used in a likelihood-based model comparison test (Brennan et al., 2016) where the surprisal, the experiment-related and stimulus-related (confounds) predictors were treated as fixed effects and the subjects as random effects. First, a specific base model containing the stimulus-related confounds predictors (LF, WD and RMS) and the experiment-related predictor (speech direction) was created for each surprisal model (two different models in order to take into account the different effects of the Gram-Schmidt procedure applied). Second, the corresponding surprisal predictors (i.e. one for the “real” and one for the “reversed” speech) were added to each base model to create a specific augmented surprisal model.

In order to compare the two augmented statistical models (i.e. the ones that include the surprisal as first predictor) in their ability to explain the variance of the neural data (i.e. the BOLD signal from each ROI), a simulated likelihood ratio test (LRT) (Royaton and Thompson, 1995) was performed. The LRT is used to compare (and choose between) two competing statistical models. For example, within the current literature of natural language processing applied to fMRI, Brennan et al. (2016) have shown how LRTs can be repeatedly applied to hierarchically compare different syntactic models, with the purpose to evaluate the unique contribution made by each type of syntactic structure to the BOLD signal across ROIs (Brennan et al., 2016). Within the context of generalized linear models, the simulated LRT can be seen as an extension of the LRT enabling the comparison of two competing statistical models that are not nested into one another (see, e.g., Lewis et al., 2011). In the present study, the two competing models used in the fMRI data analysis contain the same number of predictors, and therefore they are not nested.

In short, the simulated LRT is performed by first creating a reference (null) distribution for the LRT statistic using one of the two models (chosen as reference model). Then, a statistical significance for the other (alternative) model (given the null hypothesis that the two models are not different) is obtained non-parametrically by comparing the observed LRT statistic with the reference distribution (see, e.g., Lewis et al., 2011). In our case, the SwS was naturally chosen as the alternative model, as it was originally introduced as a special version of the LS, which was, instead, chosen as reference model. Thus, for each iteration of the simulation process, a new simulated data set was created from the LS model with the parameters set to the maximum likelihood estimates (as resulting from the fitting to the observed data). Then, both models (LS and SwS) were fitted to the simulated data set, and the result of the LRT was evaluated with respect to the reference distribution (Lewis et al., 2011). After 10,000 simulations, the estimated \( p \)-value of the model comparison was obtained as the ratio of the number of times the simulated LRT statistic is equal to, or exceeds, the observed value plus one, divided by the total number of iterations plus one (Manly, 2007).

2.10.4. Correlation between behavioral and neural responses

The regression coefficients for the surprisal predictors from the first-level GLM fit of the regional BOLD signals were extracted for each participant and each region of significant activations in the contrast “real vs. reversed speech”. These values and the percentage of correct answers of the participants in the post-hoc questionnaire were analyzed using a generalized linear model, where the regression coefficients were considered as fixed factors. The statistical significance of the effect of the regression coefficients was evaluated with an F test.

All the regional analyses were performed in MATLAB.

3. Results

Participants correctly answered at least five out of the eight questions (mean \( \pm \) std 86% ± 10%) presented in the questionnaire, indicating a satisfactory level of comprehension of the story (\( p < 0.05 \), binomial distribution, chance level: 0.25). Three subjects were excluded from the fMRI data analysis due to excessive motion or incomplete data acquisition, one subject was excluded due to a technical problem which caused the loss of the data set, resulting in a group of 27 participants (21 females).

Across included participants, translation and rotation movements were all below the threshold of 3mm and 3 degrees (\( N = 27 \), translation: mean \( \pm \) std 0.34 ± 0.40 mm; rotation: mean \( \pm \) std 0.32 ± 0.37 degree). The mean FD of each participant was below 0.5 mm (Power et al., 2012) and the grand mean FD of the analyzed cohort (\( N = 27 \), mean \( \pm \) std 0.11 ± 0.10 mm) was below the grand mean FD of an exemplary “low-motion” cohort (\( N = 43 \), mean \( \pm \) std 0.29 ± 0.12 mm) analyzed in Patel et al. (2014).

In the surprisal analysis of the narrative text, the LS model yielded a higher average surprisal compared to the SwS (2.74 vs. 2.85) with the SwS model almost systematically reducing the actual probability space for each upcoming word in comparison with the LS model. In order to test if this difference was only due to the different word contexts (and more specifically to the higher number of words) used in the estimation of the SwS model, the LS was linearly interpolated with the word conditional probabilities estimated from exactly the same context words used in the SwS estimation by creating the \( \text{LS}_{\text{sw}} \) model. In this case, the lowest average surprisal on our narrative stimulus (i.e. the best model) was obtained by choosing an interpolation factor of 0.98, indicating that the integration of the LS with further conditional probabilities estimated on the additional context used in the SwS calculation had a negligible impact on the model (Fig. 1). Therefore, as the improvement over a purely Markov model (i.e. the LS) is very small (LS average surprisal 2.86 ± 1.60, \( \text{LS}_{\text{sw}} \) average surprisal 2.86 ± 1.59), with changes in the fMRI pre-
Table 1
Clusters of statistically significant fMRI activations (p < 0.05, cluster level corrected, cluster-forming threshold p = 0.001) for both surprisal predictors. Activated regions had a higher positive correlation with the surprisal of the narrative during actual compared to reversed speech. The description of the area, the Talairach coordinates of the peak voxel, the cluster extent, and the t-value of the peak voxel in the cluster are displayed in the table.

| LS             | Talairach coord. (x,y,z) | t-value | mm³  |
|----------------|--------------------------|---------|------|
| R-STG          | 57, −11, 8               | 6.09    | 717  |
| R-ATL          | 45, 19, −20              | 7.58    | 1173 |
| R-cerebellum   | 15, −73, −40             | 5.08    | 294  |
| L-STG/MTG      | −49, 11, −18             | 8.78    | 11831|
| SwS            |                          |         |      |

| LS             | Talairach coord. (x,y,z) | t-value | mm³  |
|----------------|--------------------------|---------|------|
| R-ATL          | 45, 19, −20              | 7.17    | 1356 |
| L-STG/MTG      | −49, 11, −18             | 9.40    | 11770|
| L-IFG          | −55, 19, 20              | 6.57    | 869  |

Table 2
Summary of the results of the likelihood-based model comparisons considering the neural data of all the subjects. The SwS model yielded significant better fitting compared to the LS model in the L-STG/MTG, considering the areas where the LS showed higher activation in the original compared to the reversed speech, and in all the regions (R-ATL, L-STG/MTG and L-IFG) where the SwS yielded higher activations in the original compared to the reversed speech.

| Likelihood-based model comparisons | Area | LS vs SwS |
|-----------------------------------|------|-----------|
|                                   | R-STG| 0.99      |
|                                   | R-ATL| 0.13      |
|                                   | R-Cerebellum| 0.87 |
|                                   | L-STG/MTG| <0.05* |

| Likelihood-based model comparisons | Area | LS vs SwS |
|-----------------------------------|------|-----------|
|                                   | R-ATL| <0.05*    |
|                                   | L-STG/MTG| <0.05* |
|                                   | L-IFG| <0.05*    |

dictor on the second decimal digit, only the LS model was used in the fMRI analysis.

In the whole-brain voxel-based fMRI data analysis, LS and SwS predictors produced both overlapping and distinct activations in the contrast between original and reversed story conditions. In all significant clusters, the fMRI time-courses were found to be positively modulated by word surprisal (i.e. higher surprisal levels were associated with higher activation levels and vice versa) and the size of this effect was significantly higher when listening to the original, compared to the reversed, speech condition.

The LS predictor elicited significant activations (p < 0.05, cluster level corrected, cluster-forming threshold p = 0.001) in both hemispheres. In the left hemisphere, an activation was observed in a compact cluster extending from the posterior to the anterior portion of the superior temporal gyrus and the middle temporal gyrus (L-STG/MTG). In the right hemisphere, significant activations were detected in the superior temporal gyrus (R-STG), in the anterior temporal lobe (BA 38) (R-ATL) and in the right cerebellar lobule VIb (R-cerebellum) (Fig. 2, Table 1).

The SwS predictor activated a more left-lateralized pattern in the original, compared to the reversed, speech condition. Similar to the LS, a significant activation (p < 0.05, cluster level corrected, cluster-forming threshold p = 0.001) was observed in a cluster encompassing the L-STG and L-MTG and in a cluster in the R-ATL. These two clusters overlapped by 90% (10681 mm³) and 85% (1080 mm³) respectively between the LS and SwS models. Differently from the LS, an additional cluster was detected in the left inferior frontal gyrus (L-IFG), encompassing (part of) left pars-triangularis (BA 45) and the left pars-opercularis (BA 44) (Fig. 3, Table 1).

The subject-specific time-courses extracted from the above mentioned clusters were used in a likelihood-based model comparison to evaluate differences between models in explaining the BOLD signal and to evaluate the prediction of behavioral performances from the individual model coefficients. In all regions where the LS (or SwS) predictor elicited significantly higher activation in the real, compared to the reversed, speech condition, both surprisal models significantly improved the fitting compared to the base models without surprisal predictor. Most notably, the SwS yielded better fitting compared to the LS model in the L-STG/MTG, R-ATL and L-IFG (p < 0.05) (Table 2). No significant effects were observed for the regression coefficients of the fMRI analysis, extracted from each ROI, in the analysis of the participants’ answers to the post-hoc questionnaire (p > 0.05).

4. Discussion

In this study, the neural correlates of linguistic predictions have been investigated using naturalistic fMRI (natural listening of full-spoken nar-
tatives) and a novel semantics-weighted word-level surprisal model of continuous fMRI responses. The main objective was to augment a pure lexical surprisal model of fMRI responses by integrating lexical and semantic information within the same probabilistic language framework (Mitchell and Lapata, 2009; Sayeed et al., 2015) in the expression attempt to more selectively capture the neural responses associated with word (un)predictability in a natural listening scenario. Notably, this model has been shown to have lower perplexity on a held-out data set (Mitchell and Lapata, 2009) and to elicit better performances in predicting word pronunciation duration compared to the trigram model (Sayeed et al., 2015), although it was not previously employed for studying continuous spoken language with fMRI.

4.1. Surprisal modeling issues

The basic assumption of surprisal modeling is that a more surprising (i.e. less predictable) word stimulates greater prediction errors which may, in turn, result in more effective modulations of neural responses in those brain regions that handle such errors and contribute to successful speech comprehension (Friston et al., 2010; Tuynenhoff and Noppeney, 2016). On these premises, the neural correlates of two (LS, SwS) word-level surprisal predictors were comparatively analyzed within a mixed-effects GLM of continuous fMRI recordings. As the SwS model was expressly formulated as a weighted version of the LS model (assigning exactly the same LS values to the function words and resulting in a modulation with a semantic factor of the LS values for the content words), the two models were not combined in the first-level GLM and two separate whole-brain statistical analyses were performed. Indeed, as expected, for the narrative text used here, the SwS and LS predictors were highly (positively) correlated between each other ($r = +0.77$, see section 1 of the supplementary material for more details), thereby the SwS predictor was strictly intended to possibly replace, and not be added to, the LS predictor, in the fMRI data model. In contrast, three stimulus-related confounds were added to the first-level GLM to properly account for the variability in linguistic (lexical frequency) and non-linguistic (duration, acoustic energy) word features, that should in principle not affect the context-specific predictability of a word when this is embedded in a spoken narrative. However, because infrequent words are in general less predictable (i.e. more surprising), among context-independent linguistic features, the neural effects of the lexical frequency were also considered in a separate “context-free” whole-brain analysis using LF as predictor of interest and WD and acoustic energy as confounds (see Section 1 of the Supplementary material).

The SwS model showed a lower average surprisal value for the narrative stimulus compared to the LS, suggesting that, on average, the semantic information mostly reduces the conditional probability of a word in a context. This is apparently in contrast from what could be expected on the basis of a previous work suggesting that the integration of semantic distance with the word predictability may not be necessarily effective to improve the prediction performances (Frank and Willems, 2017). However, in our study, the context words used for estimating the lexical and the semantic components of the SwS model were purposely kept different and independent (Mitchell and Lapata, 2009; Sayeed et al., 2015) and were integrated using a scaling approach instead of a linear interpolation (Mitchell, 2010).

When comparing the raw numeric values of the LS and SwS models on content words in the narrative, there were some visible extrema in the SwS values (Fig. S2, see Section 2 of the Supplementary material for more details). However, these only occurred for 47 (out of a total of 893) content words (i.e. ~6% of the content words and ~3% of all the words in the narrative text), mostly in the lower quartiles of LS values, and no particular patterns of heteroscedasticity between the two measures were noted across LS quartiles (see Fig. S2). Moreover, when performing the same comparative analysis on the two resulting fMRI predictors (i.e. the series of values across time points of the fMRI series, not across words of the narrative, after preprocessing and hemodynamic convolution), the resulting scatter plot showed no extrema (and no heteroscedasticity patterns) for the SwS predictor (see Fig. S3).

Finally, to test if the observed difference between the two models was just due to the inclusion of a larger word context in the SwS, a different lexical surprisal model was estimated by following the workflow described in Frank and Willems (2017). In present study, the further interpolation of the LS with a factor that takes into account the conditional probabilities derived from the broader (non-local) word context used to estimate the semantic component of the SwS model, had very little impact on the LS model, and this observation is in agreement with the analysis performed by Frank and Willems (2017) on the English language (Frank and Willems, 2017).

4.2. Surprisal-related neural effects

Both the LS and SwS predictors elicited a significant activation in the L-ATG/MTG and in the R-ATL in the fMRI contrast between the actual speech (real story) and the reversed speech (reversed story). However, while the LS also activated the R-ATG and the R-cerebellum, the SwS selectively activated a cluster in the L-IFG. The activation of this network of areas is perfectly in line with a recent work by Shain et al. (2020) which also showed how the cognitive processes underlying linguistic prediction, as indexed by word co-occurrence models...
such as the surprisal models, are fully supported by the language network and not by more general domain areas (Shain et al., 2020).

Notably, the more extended activation pattern obtained via the SwS model, encompassing both the whole L-STG/MTG and the L-IFG, could not be equally obtained by the sole semantic component (i.e. the semantic scaling factor isolated from the SwS formulation). Instead, a separate whole-brain fMRI analysis based on the semantic component as the predictor of interest (see Section 2 of the supplementary material for more details) revealed that, when isolating the semantic scaling factor from the SwS model, and using it to modulate an independent word-level measure of semantic relatedness of each content word with its previous context, the significant effects of the resulting fMRI contrast within the language network were substantially reduced and mainly confined within the secondary auditory cortex (see Fig. S5). Thus, considering also that the semantics weighting mainly produced a linear ($R^2=0.78$), rather than exponential ($R^2=0.08$), transform of the semantic similarity values across the content words (i.e. changes in semantic similarity and semantic weighting, after the transformation and normalization, were mainly proportional, see Fig. S4), it is plausible that the increased sensitivity of the SwS model within the language network may truly reflect the constructive integration of a context-dependent semantic similarity with the conditional probabilities.

On the other hand, a context-free whole-brain analysis (i.e. with the LF as predictor of interest) also revealed significant positive effects (i.e. more infrequent words were associated with higher positive BOLD signal changes) within the same language network, albeit the shape and size of the cortical patches were different from the corresponding ones highlighted by the LS and SwS predictors (Fig. S1, see Section 1 of the supplementary material for more details). Thus, similar to previous studies (Levy, 2008; Shain et al., 2020; Staub, 2015), we also observed that, both LF and (two) surprisal measures can capture a significant portion of the BOLD signal variance within the language network. These observations further support the notions that lower-frequency words are also (generally speaking) more unexpected than higher-frequency words (Staub, 2015) and that, during natural listening, word frequency effects are strictly linked with word predictability effects at the neural level (Levy, 2008; Shain et al., 2020).

It should be noted also, that the sign of neural effects for frequency was not always consistent among previous studies with the (most obvious) expectation that more infrequent words trigger higher (positive) BOLD signal changes (implying that the processing of these words should be carried out with greater neural processing cost). Actually, while the sign of frequency-related neural effects was positive in our naturalistic fMRI data, in full agreement with the previous work by Staub (2015), the recent study by Shain et al. (2020) found exactly the opposite, i.e. more infrequent words were associated with lower BOLD signal changes. Thus, we could speculate that the impact of the LF measure, albeit potentially highly significant, could be in principle less specific in the prediction of the BOLD signal changes within the language network, possibly due to its strict dependency on the specific corpus with no link to the local context of the narrative being listened during measurements. However, future studies are needed to clarify the neural interactions between lexical surprisal and lexical frequency within the language network, as also suggested by Shain et al. (2020).

4.3. Whole-brain analysis of surprisal-related fMRI activity

The STG and the MTG are two essential cortical structures for both speech comprehension (Binder et al., 2009, 2008; Friederici et al., 2003; Lopopolo et al., 2017) and voice recognition (Belin et al., 2011; Perrett et al., 2015; Watson et al., 2014).

The bilateral involvement of the STG from the LS predictor applied to an Italian story is perfectly in line with a previous study using an identical implementation of the lexical surprisal model applied to Dutch stories. Thus, this finding also proposes the postulated prediction mechanism whereby the STG activity is modulated by the prediction (error) of the upcoming lexical form of a word (Willems et al., 2016). Contrariwise, the lack of activation of the R-STG for the SwS model might be due to the fact that the lexical predictability would involve the right hemisphere more heavily (Bonhage et al., 2015; Carter et al., 2019) whereas the semantic prediction (that is strictly related to the semantic system Carter et al., 2019) should be more left-lateralized (Binder et al., 2009; Carter et al., 2019).

For what concerns the involvement of the L-MTG, it has been previously suggested that this region is deeply involved in speech processing, and, more specifically, in triggering the retrieval of lexical and syntactic features of an incoming word from the long-term memory (Dronkers et al., 2004; Lopopolo et al., 2017). However, it also plays a fundamental role in facilitating the collection and storage of semantic knowledge related to the words (Binder et al., 2009). Therefore, the activation of this cortical region for both LS and SwS predictors (but also the better fitting found for the SwS predictor) would be in line with the interpretations provided in two previous studies using models of linguistic prediction and reporting the link between the L-MTG activation and both lexical (Lopopolo et al., 2017) and semantic (Weber et al., 2016) predictability of the stimulus.

The activation detected in the right posterior lobe of the cerebellum corroborates the hypothesis, supported by previous findings, that this area has a role in the linguistic expectancy (Argyropoulos, 2016; Moerger and Ivy, 2016; Pleger and Timmann, 2018; Sokolov et al., 2017). Particularly, the cerebellum has been found to be responsive to the predictability of the upcoming word (Lesage et al., 2017), to the prediction of semantic (Carter et al., 2019; D’Mello et al., 2017) and syntactic (Carter et al., 2019) features, and, more in general, during several language processing tasks (Marién et al., 2014).

The bilateral activation of the ATL has been observed by Willems et al. (2016) using an identical implementation of the LS on a different language (Dutch) (Ferstl et al., 2008; Willems et al., 2016). Our finding replicates the same mechanism in Italian and therefore both strengthen the notion of this region implicated in language processing in general and, more specifically, in the integration of linguistic information within a stream of words (Carter et al., 2019) thereby confirming the sensitivity of the surprisal measure in capturing the modulation of neural responses underlying the comprehension of the text.

The left IFG has been extensively associated with language-related processing tasks such as syntax computation, semantic selection and phonological demand tasks (Ardila et al., 2016; Binder et al., 2008; Carreiras et al., 2012; Clos et al., 2014). However, studies focusing on linguistic prediction showed mixed findings (Henderson et al., 2016; Lopopolo et al., 2017; Willems et al., 2016). In our study, the SwS elicited a significant activation in the L-IFG encompassing both the pars opercularis and the pars triangularis, which seems in line with the putative role of the L-IFG in the semantic prediction as a higher-level cognitive control processing unit (Binder et al., 2009; Ferstl et al., 2008; Hagoort, 2013, 2005; Thompson-schill et al., 2009; Zhu et al., 2012). The activation of the L-IFG for the SwS predictor and the better fitting showed by the model, therefore, suggest that the semantic integration gathers a better (more sensitive) measure of the ongoing high-level cognitive engagement required by the linguistic prediction than the LS predictor (Mitchell and Lapata, 2009; Sayeed et al., 2015).

The lack of activation in the left fusiform gyrus in both LS and SwS patterns is partly surprising, as its activation has been previously reported using two different measures of linguistic prediction, albeit on the same data set (Lopopolo et al., 2017; Willems et al., 2016). However, two different explanations were provided in those studies: Willems et al. (2016) pointed out the emergence of priming effects in the fusiform gyrus which were successfully detected by the model because of the prediction (error) in the word form (Willems et al., 2016). Instead, Lopopolo et al. ascribed this effect to the continuous access to the lexico-semantic information of the word via this region (Lopopolo et al., 2017). In our case, we might speculate that the activation of this region did not reach the statistical significance for both the LS and the SwS models.
most likely because, compared to Dutch and English, Italian has higher orthographic transparency and lower syllable complexity (Borleffs et al., 2017; Seymour et al., 2003) whereby the postulated continuous access to the word form might be less necessary in general and consequently the surprisal modulation would be less effective.

4.4. Region-based comparison of surprisal models of fMRI activity

The fMRI time-courses from the above clusters were also used to compare the ability of the surprisal predictors to fit these neural data. Particularly, the use of a simulated LRT allowed a rigorous comparison of the two statistical models that, although related to one another, could not be nested as they included the same number of predictors and the first predictor was an alternative version of the surprisal model.

The comparisons of the two surprisal models with their corresponding base models confirmed that the use of a surprisal predictor may help in explaining more variance of the neural response as measured with fMRI during natural listening. Moreover, in comparison with the LS, the SwS model showed a significantly better performance in fitting the neural data in the L-STG/MTG, R-ATL and L-IFG. Altogether these findings support, first, the idea of using surprisal models to capture more variance in the BOLD response during language comprehension and, second, the hypothesis that the integration of the semantic information into the surprisal probabilistic framework may further improve the ability of the model in fitting the neural responses.

4.5. Behavioral correlates of surprisal-related fMRI activity

The regression coefficients from the ROI analysis of fMRI time-courses in the above clusters were also used in the analysis of the participants’ answers to the post hoc questionnaire to test the possible correlation between the magnitude of the BOLD response and the participants’ level of comprehension. In fact, as the surprisal measure has been proven to be a good proxy of the cognitive effort related to linguistic processing (Demberg and Keller, 2008; Hale, 2016, 2001; Levy, 2008; Smith and Levy, 2013) it was possible that the higher (lower) the cognitive effort required for integrating the incoming input with the current information, the more positively (negatively) the BOLD signal had to be expected to change within the activated regions. As a consequence, the variation of the BOLD response of a participant could signal a higher (lower) cognitive engagement in the comprehension of the entire sentence albeit this does not necessarily imply that the same participant is ultimately understanding more (less) of its meaning.

The analysis of the participants’ answers did not yield a significant effect of the regression coefficients of the group fMRI analysis on the participants’ scores, therefore we cannot provide supporting evidence to the possibility of using fMRI measures as a possible index of comprehension. On the other hand, as all participants were healthy, fully compliant and highly educated young subjects, most of them responded correctly to the majority of the questions, and therefore it is likely that the observed variability in the participants’ scores was far too small to reveal possible correlations with the BOLD responses. Thus, we cannot rule out (and actually would rather still expect) that, in future studies, the use of more complex stories and/or the submission of far more detailed questionnaires (e.g. more questions, more choices per question, etc.) or the enrollment of subjects overtly experiencing more difficulties in understanding the narrative (e.g. hypoacusic or mentally retarded patients) could gather more inter-subject variability in the participants’ levels of understanding that can be better correlated with the size of surprisal-related BOLD effects.

4.6. Future directions

In general, the proposed experimental paradigm has a number of advantages that could make it useful or convenient for certain clinical applications. In fact, the use of a naturalistic stimulus (i) enables the possibility to study language-related brain mechanisms without involving the subjects in an explicit cognitive performance and (ii) may improve the overall quality of the data as naturalistic fMRI measurements have been shown to be more accurate and reproducible than resting-state fMRI measurements and also less vulnerable to the major confounds of head motion and physiological artifacts (Chen et al., 2019; Hasson et al., 2010). Particularly, the possibility to have informative neural predictors that target one or more linguistic aspects from the BOLD response changes during listening to a spoken narrative could be clinically relevant for studying subjects affected by receptive language disorder (i.e. a syndrome characterized by problems in language understanding but not production, mostly present in the children and adolescents (Goldstein and Naglieri, 2011)), as it can be used to obtain a fixed response model for both a healthy and a pathological group without the need of co-varying the response with the behavioral performance. For instance, in a cohort of participants affected by language comprehension deficits, it is likely to expect different neural behavior in key areas of the language network compared to the pattern elicited in a population with normal language abilities. However, as this approach has never been used (to the best of our knowledge) in this scenario we cannot yet formulate clear expectations about the direction of these changes.

Finally, the existence of natural language processing models based on recurrent neural networks, that appear to learn both the semantic features and the position of words within the streams (Bengio et al., 2003; De Mulder et al., 2015), suggests that future investigations and developments of the presented surprisal models, and in particular the SwS model, should possibly attempt to take the position of the words in the context into account.

5. Conclusions

A novel variant of the surprisal model for word prediction incorporating both lexical and semantic information has been presented and used to map the putative neural correlates of a story comprehension in a naturalistic fMRI experiment. Compared to a pure lexical surprisal, the semantics-weighted surprisal provides a better predictor of the fMRI activity in left STG and MTG and is critically more sensitive in detecting linguistic processes in the left IFG. While future studies are certainly warranted, as far as the amount of surprisal modulation detectable from continuous fMRI recordings can be interpreted as an index of the cognitive engagement during language comprehension in subjects listening to spoken narratives, the proposed approach may gather a possible neurological marker of (in)sufficient understanding of a story. In this case, analyzing different groups of people (e.g. from healthy and clinical populations) listening to the same story (or to different stories of varying language and complexity), eventually under (physiologically or pharmacologically) modified conditions, will possibly expand or challenge the validity of the proposed models. Finally, the use of highly naturalistic stimuli such as full-spoken narratives could be a crucial feature for studying linguistic processes in poorly compliant patients such as children or elderly patients.

Data Sharing Statement

Individual de-identified participant imaging data in NIfTI format, de-identified pre-processed data, code and scripts are available at https://github.com/andreagrusso/fMRIandSurprisalItalian

Informed consent

All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1975, and the applicable revisions at the time of the investigation. Informed consent was obtained from all patients for being included in the study.
Declaration of Competing Interest

The authors report no disclosures.

CRediT authorship contribution statement

Andrea G. Russo: Conceptualization, Methodology, Software, Investigation, Formal analysis, Writing - original draft, Data curation, Visualization. Maria De Martino: Writing - review & editing, Project administration. Azzurra Mancuso: Project administration. Giorgio Iaconetta: Writing - review & editing. Renzo Manara: Writing - review & editing. Annibale Elia: Supervision, Resources. Alessandro Laudanna: Supervision, Resources. Francesco Di Salle: Supervision. Fabrizio Esposito: Conceptualization, Writing - review & editing, Visualization, Supervision, Project administration.

Funding resource

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgments

The authors thank Anna Scibelli for the useful discussions about the syntax and the morphology of the Italian language.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2020.117281.

References

Anderson, J.J.R., Skare, S., Ashburner, J., 2003. How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. Neuroimage 20, 870–888. doi:10.1016/S1053-8119(03)00367-7.

Aréda, A., Bernal, B., Roselli, M., 2016. How localized are language brain areas? A review of broadman areas involvement in oral language. Arch. Clin. Neuropsychol. 31, 112–122. doi:10.1016/j.acnnb.2009.01.021.

Argyropoulos, G.P.D., 2016. The cerebellum, internal models and prediction in ‘non-motor’ aspects of language: a critical review. Brain Lang.: Contrib. Cerebellum Lang. Funct. 161, 4–17. doi:10.1016/j.bandl.2015.08.003.

Armeni, K., Willems, R.M., Frank, S.L., 2017. Probabilistic language models in cog- nitive neuroscience: promises and pitfalls. Neurosci. Biobehav. Rev. 83, 579–588. doi:10.1016/j.neubiorev.2017.09.001.

Belin, P., Bestelmeyer, P.E.G., Latimé, M., Watson, R., 2011. Understanding voice perception. Br. J. Psychol. 102, 711–725. doi:10.1111/j.2044-8295.2011.02041.x.

Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C., 2003. A Neural Probabilistic Language Model. J. Mach. Learn. Res. 3, 1137–1155.

Binder, J.R., Desai, R.H., Graves, W.W., Conant, L.L., 2009. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cereb. Cortex 19, 2767–2796. doi:10.1093/cercor/bhp055.

Binder, J.R., Swanson, S.J., Hammeke, T.A., Sabsevitz, D.S., 2008. A comparison of five fMRI protocols for mapping speech comprehension systems. Epilepsia 49, 1980–1997. doi:10.1111/j.1528-1167.2008.01618.x.

Bird, S., Lopez, E., Klein, E., 2009. Elaboration del Lenguaje Natural con Python. Bonhage, C.E., Mueller, J.L., Frueridi, A.D., Fiebach, C.J., 2015. Combined eye tracking and fMRI reveals neural basis of linguistic predictions during sentence comprehen- sion. Cortex 68, 33–47. doi:10.1016/j.cortex.2015.04.011. Special Issue: Prediction in speech and language processing[http://doi.org/].

Borlffé, E., Bannen, A.M., Lyttinen, H., Zwarts, F., 2017. Measuring orthographic transparency and morphological-syntactic complexity in aphasic orthographies: a narrative review. Read. Writ. doi:10.1111/1444-012X.12392.

Brennan, J.R., Stabler, E.P., Van Wagenen, S.E., Luh, W.M., Hale, J.T., 2016. Abstract lin- guistic structure correlates with temporal activity during naturalistic comprehension. Brain Lang. 157, 158–194. doi:10.1016/j.bandl.2016.04.008.

Brown, E.C., Muzik, O., Rothermel, R., Matsuzaki, N., Juhász, C., Shak, A.K., Atkin- son, M.D., Fuerst, D., Mittal, S., Sood, S., Divadkar, V.A., Asano, E., 2012. Evaluating reverse speech as a control task with language-related gamma activity on electrocere- tigraphy. NeuroImage 60, 2335–2345. doi:10.1016/j.neuroimage.2012.02.040.

Carreiras, M., Pattamadilok, C., Meseguer, E., Barber, H., Devlin, J.T., 2012. Broca’s area plays a causal role in morphosyntactic processing. Neuropsychologia 50, 816–820. doi:10.1016/j.neuropsychologia.2012.01.016.

Carter, B.T., Foster, B., Muney, N., Luke, S.G., 2019. Linguistic networks associated with lexical, semantic and syntactic predictability in reading: A fixation-related fMRI study. Neuroimage 189, 224–240. doi:10.1016/j.neuroimage.2019.01.018.
