Words in context: tracking context-processing during language comprehension using computational language models and MEG

Alessandro Lopopolo^{1*}, Jan Mathijs Schoffelen², Antal van den Bosch^{3,4}, Roel M. Willems^{1,2,5}

1 Centre for Language Studies, Radboud University Nijmegen, Nijmegen, the Netherlands

2 Donders Institute, Radboud University Nijmegen, Nijmegen, the Netherlands

3 Meertens Institute, Royal Netherlands Academy of Science and Arts, Amsterdam, the Netherlands

4 University of Amsterdam, Amsterdam, the Netherlands

5 Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands

* Corresponding author: a.lopopolo@let.ru.nl

Abstract

The meaning of a word depends on its lexical semantics and on the context in which it is embedded. At the basis of this lays the distinction between lexical retrieval and integration, two basic operations supporting language comprehension. In this paper, we investigate how lexical retrieval and integration are implemented in the brain by comparing MEG activity to word representations generated by computational language models. We test both non-contextualized embeddings, representing words independently from their context, and contextualized embeddings, which instead integrate contextual information in their representations. Using representational similarity analysis over cortical regions and over time, we observed that brain activity in the left anterior temporal pole and inferior frontal regions shows higher similarity with contextualized word embeddings compared to non-contextualized embeddings, between 300 and 500 ms after word presentation. On the other hand, non-contextualized word embeddings show higher similarity with brain activity in the left lateral and anterior temporal lobe at earlier latencies – areas and latencies related to lexical retrieval. Our results highlight how lexical retrieval and context integration can be tracked in the brain using word embeddings obtained with computational models. These results also suggest that the distinction between lexical retrieval and integration might be framed in terms of context-independent and contextualized representations.

Introduction

Sentence comprehension is thought to rely on two basic operations, lexical retrieval and integration [1–6]. Lexical retrieval refers to the process that extracts the meaning of individual words from the mental lexicon [1,2,6], which takes place before an integration process combines the individual semantic representations in meaningful ways.

Baggio & Hagoort [7] argue that it is difficult to directly test the nature of the distinction between retrieval and integration using traditional task-based experimental paradigms, since it might be impossible to devise a task tackling only one operation while leaving the other untouched. Therefore, in the study presented here, we instead propose to use computational semantic models to further investigate the neural basis of

these two basic operations.

In the present study, we use computational semantic models to investigate the neural basis of lexical retrieval and integration. We use two classes of computational semantic models, representing linguistic items either as independent or dependent from their context of occurrence. In the computational linguistics literature, these models are usually referred to as non-contextualized and contextualized embeddings, respectively [8,9]. By comparing them to MEG data collected during sentence comprehension, we aim to get more insight into the neural basis of these processes by showing that integration is approximated by contextualized embeddings and that it is a separate process (both functionally and physiologically) from semantic memory. The main methodological contribution of computational modeling to the study of language processing in the brain relies on the fact that it helps in avoiding the limitations of task-oriented studies by exploiting the richness of naturally occurring sentences and by relying on a fleshed out model of the investigated processes. In other words, computational linguistic modeling provides a more direct implementation of the process and does not require the assumption that the process can be decomposed in orthogonal subprocesses that can be controlled by specific experimental tasks.

Integration as contextualization

The distinction between retrieval and integration can be grounded on the observation that the language system seems able to deal with a virtually infinite number of utterances, which in turn seem to be composed by a limited, although flexible, set of primitives, such as phonemes, morphemes, words, and phrases. By storing these primitives in a hypothetical partition of long-term memory, recalling them when necessary, and combining them in a seemingly unbound number of configurations, humans can deal with a large variety of messages in a parsimonious, flexible and creative way.

The present study frames the concept of integration in terms of contextualization of lexical items. This is based on the assumption that the meaning of a word depends on how it is represented in memory, and how this representation changes as a result of its integration in context, linguistic in our case, it is presented in.

(1) In order to open a new account, you should go to a **bank**.

(2) A fisherman is sitting with his rod on the **bank** of the river Thames.

For instance, humans distinguish the meaning of *bank* as “building or financial institution” or as “the shore of a river” depending on whether it is encountered in the context of Sentence 1 or 2. The presence of the string “*a new account*” in Sentence 1 steers the interpretation towards the financial domain, whereas the string “*a fisherman*” acts as bias towards a river-related interpretation of the word **bank**.

A distinct yet similar disambiguation problem arises with tokens that are not lexically ambiguous, such as *dog* in Sentences 3 and 4. Although the basic meaning of the term is the same in the two sentences, the two tokens take on two distinct sentential roles: as the object of an action (4) or the subject of a statement realized as a compound nominal predicate (3).

(3) The domestic **dog** is a member of the genus *Canis*, which forms part of the wolf-like canids.

(4) I took my **dog** out for a walk in the park.

It has been suggested that the human brain creates representations of words that are different according to such contextual cues [10].

Recent years have seen an increase of interest in the role of context in the construction of artificial representations of word meanings in the field of computational linguistics. Striking results have been reported with so-called contextualized word embedding models, showing significant improvements in performance in several computational linguistic tasks as compared to those obtained using word embeddings (representations) that do not consider context [9, 11–14]. In this study, we investigated the effect of contextualization of words on brain activity, comparing non-contextualized and contextualized models. Our working hypothesis is that, after lexical retrieval from memory, brain activity can be modelled by non-contextualized embeddings. Conversely, the product of lexical integration can be modelled by contextualized embeddings.

Brain activity related to language comprehension reflects processes that involve different areas of the brain at different moments in time following the onset of the

stimulus [15–17]. It is therefore capital to show that the putative similarity between a model and a brain process concerns not only areas associated with such process, but also that it does so in a time frame that is compatible with the time course of language processing. For this reason, we use a magnetoencephalographic (MEG) dataset collected during sentence reading. MEG records brain activity at the level of milliseconds, and with a reasonable anatomical resolution, making it ideal for a study interested also in the when, and not only the where of a specific neural process [18].

Lexical processing in the brain

Neural loci

Semantic memory, a component of long-term memory, acts as the storage of knowledge and representations of basic linguistic units, such as words. In a simplified manner, we can define memory as the mental lexicon, or the equivalent of a vocabulary in which the representations of words are stored and wait to be retrieved during production or comprehension. Binder and colleagues [19,20] provide a comprehensive picture of the cortical areas substantiating semantic memory. Memory, semantic memory in particular, is associated with the lateral portion of the left temporal cortex (middle temporal gyrus), parts of the inferior temporal gyrus, and the inferior parietal cortex. An important role is also hypothesized for the anterior portions of the temporal lobe (anterior temporal pole, ATP). The involvement of the ATP is confirmed by both studies on semantic dementia [21,22], and by a large body of neuroimaging literature [23–26]. These findings have been summarized by Patterson & al. 2007 [27] and led to the formulation of the hub and spoke model, which posits that concepts are represented by a network of sensorimotor representations converging in the ATP, which acts as a hub collecting and controlling modality-specific features in order to produce supra-modal representations.

Integration is a process that operates on representations retrieved from semantic memory. In its most basic formulation, integration consists of merging two linguistic tokens (e.g., two words) and creating a larger unit, such as a phrase or, more simply, a bi-gram. Integration is an operation that takes a token and embeds it into the context

represented, for instance, by the other tokens making up the sentence in which it is presented. Brain imaging and brain lesion studies suggest that the inferior frontal gyrus, in interaction with areas in the perisylvian and temporal cortex, plays an essential role in lexical integration [17,28].

Integration also involves anterior temporal areas. By contrasting the activity recorded during the reading of sentences and of word lists, works such as Mazoyer & al. (1993) [29], Stowe & al. (1998) [30], Friederici & al. (2000) [15], Humphries & al. (2006) [31], and Humphries & al. (2007) [32] reported an increase in activity in ATP for the former condition as compared to the latter. The role of ATP in processing integration is confirmed by another series of studies narrowing down the scope of the analysis. Rather than working with sentences as a whole, these analyses focused on simple phrasal processing, consisting of the composition of a wide range of phrasal and syntactic compositional types and cross-language and modality [33–37].

Timing of processes

Besides the cortical loci of processing, sentence processing is characterized by a specific temporal profile that describes the timing of each of its sub-processes [16,32]. The earlier stages mainly concern the recognition of the word from its auditory (for spoken words) or graphic (for written words) form and involve primary auditory or visual areas between the onset of a word and 150–200 ms. The phases that interest our analysis are the so-called Phase 1 and Phase 2, as described by Friederici [16].

Phase 1 takes places after the word form has been identified, and can be broken down into sub-phases. First, around 200 ms after the onset of a word, its category is identified (i.e., whether it is likely to be a verb, a noun, an article, etc.). Subsequently, information of its lexical meaning is retrieved from semantic memory implemented in the middle temporal gyrus. This process takes place approximately between 150 and 300 ms after the onset of a word.

Phase 2 follows between 300 and 500 ms after stimulus onset. It roughly corresponds to integration, as introduced in the previous Section. In this phase, the lexical representation of a word retrieved in Phase 1 is embedded in the contextual representation consisting of the retrieved and unified representation of the other words

composing the sentence that is processed [38–41].

Computational models

For the purpose of this study, we use two types of computational models developed for generating word representations: non-contextualized models and contextualized models. These models create so-called word embeddings, which consist of vectors of real numbers populating a high-dimensional space. In other words, a model M takes a word w and returns a real vector $\vec{w}$ representing w in a high-dimensional space S .

The first type of model generates representations $\vec{w}$ that are independent from the context (sentence, paragraph, etc.) in which the represented word w is located. We call this type of model non-contextualized, and it is represented by the popular word2vec model (Section 4) [8].

Besides non-contextualized models, we also consider a contextualized model: ELMo (Sections 4 [9]. This model, contrary to word2vec, assigns representations $\vec{w}$ that depend on the textual context in which the represented word w is located. For instance, if the word *dog* appearing in Sentences 3 and 4 always obtains the same embedding $\vec{dog}^{w2v}$ from word2vec, it will obtain two different vectors ($\vec{dog}_{@s3}^{ELMo}$ and $\vec{dog}_{@s4}^{ELMo}$), one for each of the two contexts in which it is found.

As shown in Figure 1, this becomes evident when we compute the similarity between the embeddings. The cosine similarity between the word2vec generated $\vec{dog}_{@s3}^{w2v}$ vector (Sentence 3) and $\vec{dog}_{@s4}^{w2v}$ (Sentence 4) is 1.0, indicating a total identity between the two representations. Instead the cosine similarity between the ELMo generated $\vec{dog}_{@s3}^{ELMo}$ vector (Sentence 3) and $\vec{dog}_{@s4}^{ELMo}$ (Sentence 4) is 0.92. Similarly the similarity between the word2vec representations of *bank* (Sentences 1 and 2) is 1.0, whereas the similarity between the ELMo representation of the same words is 0.87.

Fig 1. The effect of contextualization on pairwise word similarity.

Word2vec (blue) returns identical representations for words independently from the sentences they are presented in. This is evident from the 1.0 cosine similarity between *bank* in Sentences 1 and 2 – on the one hand – and between *dog* in Sentences 3 and 4. Conversely ELMo (red) instead produces different contextualized representations of the same words depending on the context they are used, note the lower similarities (0.92 and 0.87).

Non-contextualized embeddings (Word2vec)

Word2vec is an artificial neural network-based model used to produce word embeddings. It has been proposed as a more advanced alternative to earlier distributional semantic vector spaces such as latent semantic analysis. From an architectural point of view, word2vec consists of a shallow, two-layer neural network. The model can be trained either to predict the current word from a window of surrounding context words (continuous bag-of-words, CBOW), or – conversely – predict the surrounding window of context words given a target word (continuous skip-gram, CSG).

Training creates a high-dimensional vector space populated by word vectors, which are positioned in the space in such a way that words that share similar semantic and syntactic properties lay close to one another. Fig 2 represents the architecture of the CBOW version of word2vec which is used in the present study (adapted from [42]). The trained embeddings correspond to the weights stored in matrix W , whose dimensions $[v \times d]$ correspond to the size of the modeled vocabulary (v) and the chosen number of dimensions (d) of the vector space itself. Therefore, the way word2vec assigns embeddings to word w can be seen as a sort of dictionary “look-up” where the embedding of word w_i corresponds to row i of matrix W .

Fig 2. Architecture of the non-contextualized word embedding model (word2vec). The model embeddings (W) are trained by predicting the current word from a window of surrounding context words (continuous bag-of-words, CBOW).

For the purpose of the present study, it is important to point out the role of context with regard to the way word2vec is trained and used to assign word embeddings. Context indeed plays a capital role during the training of the model. In both cases, the context of a word is present in the pipeline, either as the input or as the target of the training function. Nonetheless, once the model is trained, its application is blind to the context and relations that the words have.

Contextualized word embeddings (ELMo)

The contextualized word embedding model ELMo [9] relies on the properties of recurrent neural networks. Contrary to word2vec, ELMo is a deep contextualized model for the generation of the word representation. It models complex characteristics of word

use that vary across linguistic contexts.

From an architectural point of view, ELMo is a recurrent bi-directional language model (biLM) composed of 2 layers of bi-directional recurrent units (implemented as LSTM layers) feeding to an output softmax layer. A language model refers to a system (stochastic or neural) trained on predicting a word given its preceding context. In its most common formulation, a language model is an approximation of the function $P(w_i|w_{1:i-1})$, i.e. a function that assigned a probability to a word w_i given its prefix $w_{1:i-1}$.

A recurrent layer is a layer that creates a representation of the input sequence (sentence) at word w_i as a combination of the representation of the i_{th} word and the representation of its preceding (if the recurrent neural network proceeds left to right) or following (if it proceeds right to left) context.

Fig 3 illustrates the structure of the ELMo model consisting of a 2-layer biLM. ELMo's bidirectionality refers to the fact that its hidden layers receive information about both the preceding and following words of w_i . The representation ELMo generates for a word consists of a concatenation of the activation of its two recurrent layers.

Fig 3. Architecture of the contextualized word embedding model (ELMo). ELMo architecture consisting of a 2-layer biLM, where the layers are implemented as LSTMs. The contextualized embedding is produced as a linear combination of its components.

Contrary to how word2vec assigns embeddings to words, ELMo does not employ a word-vector dictionary (as word2vec W in Fig 2). Given w_i in a sentence $S = [w_1, w_2, \dots, w_n]$ ELMo instead creates the embedding $\vec{w}_i$ by passing the whole S text through the biLM. Embedding $\vec{w}_i$ will depend on context $w_{1:i-1}$ and $w_{i+1:n}$ as the combination of the activation of the LSTM layers corresponding to the presentation of w_i to the model. For these reasons, if w_i appears in a different sentential context S' , its $\vec{w}_i'$ will be different.

Training

Since the MEG data (described in Section 4) was collected from Dutch native speakers reading Dutch sentences, we used models trained on Dutch texts. Word2Vec was

trained on the CBOW task on the Corpus of Spoken Dutch (CGN2). For ELMo we used the pretrained Dutch model provided by the ELMoForManyLangs collection¹ with the same hyperparameter settings as [14] and trained on a set of 20-million-words Dutch corpus randomly sampled from the raw text released by the CoNLL 2018 shared task on Universal Dependencies Parsing.

Hypotheses

This study aims at investigating the neurobiological underpinnings of the distinction between non-contextualized and contextual representation as derived from recent developments in word embedding modeling techniques. More specifically, for the **non-contextual model**, we hypothesize a similarity between this model and the operation of **lexical retrieval** from memory. We expect that a model with such characteristics should most closely resemble the activity in lateral temporal regions during processing, around 200 ms after word onset.

Conversely, **contextual model** representations of words are expected to resemble more the activity related to **integration** and described as pertaining inferior frontal and antero-temporal regions between 300 and 500 ms after the onset of the word.

Fig 4 schematically represents these hypotheses.

Fig 4. Hypotheses. The hypothesis presented in this paper is that brain activity relative to lexical retrieval can be modelled by non-contextualized embeddings, whereas that of integration can be modelled by contextualized embeddings instead.

MEG Data

We used the MEG data belonging to the MOUS dataset [18] collected at the Donders Centre for Cognitive Neuroimaging in Nijmegen, The Netherlands. For more details on the acquisition procedure, stimuli, pre-processing, and source reconstruction techniques, we refer to the original paper and to Schoffelen & al. 2018 [18, 45].

¹<https://github.com/HIT-SCIR/ELMoForManyLangs> [43, 44]

MEG data acquisition and pre-processing

The data were collected with a 275 axial gradiometer system (CTF). The signals were digitized at a sampling frequency of 1200 Hz (the cutoff frequency of the analog anti-aliasing low pass filter was 300 Hz). Head position with regards to the sensors was determined by 3 coils attached to the participant's head. Electro-cardiogram and the horizontal and vertical electro-oculogram were measured by 3 bipolar Ag/AgCl electrode pairs.

Electrocardiogram artifacts were identified based on their topography and subtracted from the data. The data was segmented into trials corresponding to activity recorded from -183 ms before word presentation to a variable time after word presentation, depending on word length. Trials that contained artifacts (Eye movements and muscle contractions and jump artifacts in the SQUIDS) were excluded from further analysis. Next, the power line interference was estimated and subtracted from the data. The data were down-sampled to a sampling frequency of 300 Hz.

Source reconstruction was performed using a linearly constrained minimum variance beamformer (LCMV) [46], estimating a spatial filter at 8,196 locations of the subject-specific reconstructed midcortical surface. The dimensionality of the data was reduced by applying an atlas-based parcellation scheme based on a refined parcellation Conte69 atlas (191 parcels per hemisphere). After that, spatial filters were concatenated across vertices comprising a parcel. The first two spatial components were selected for each parcel. For more details on this procedure we refer to Schoffelen & al. 2018 [45].

Subjects

We used the data of 74 subjects belonging to the MEG section of the MOUS dataset [18]. All subjects were Dutch native speakers, who were asked to silently read 120 Dutch sentences presented on a screen word by word, containing a total of 1377 words. All sentences varied between 9 and 15 words in length.

Stimulation paradigm

The sentences were presented visually with an LCD projector, with a vertical refresh rate of 60Hz situated outside the MEG scanning room, and projected via mirrors onto

the screen inside the measurement room. All stimuli were presented in a black
mono-spaced font on a gray background at the center of the screen within a visual angle
of 4 degrees. Sentences were presented word-by-word with a mean duration of 351 ms
for each word (minimum of 300ms and maximum of 1400ms, depending on word length).
The duration of the stimuli was determined taking into account both the word length,
the number of words in the whole sentence, and the number of letters within each word.
Each word was separated by an empty screen for 300 ms before the onset of the next
word.

Analysis

The computational models introduced in Section 4 are used to generate vectorial
representations of the same stimulus sentences presented during the acquisition of the
MEG data described in Section 4. These representations are generated at the word level,
meaning that our models assign a set of real number vectors for each word of the
stimulus sentences. The non-contextualized word embedding (word2vec) assigns only
one vector per word, whereas the contextualized models (ELMo) represent every word
with a set of vectors, each one corresponding with one component (layer) of their
internal architecture.

In our analyses, a word represented by ELMo is assigned one vectorial representation
corresponding to the average of the activation of the two bidirectional layers composing
the ELMo network ($EL\vec{M}o$).

This section describes the analysis methodologies employed to map the two vectorial
representations ($W\vec{2}V$, $EL\vec{M}o$) to the corresponding brain activity recorded with MEG.
Our goal is to both map the overlap between model representations and brain activity
at the anatomical level, and to track the temporal evolution of such similarity. For this
reason, we adopted a version of representational similarity analysis (RSA) implemented
in a way to return both spatially (anatomical) and temporally situated similarity scores.

Representational similarity analysis through time

Given a set of linguistic units, for instance, words $w_1, \dots, w_n$, we can generate a vectorial
representation for each of them using the embedding models described in Section 4.

Words $w_1, \dots, w_n$ are assigned vectors $\vec{w}_1^M, \dots, \vec{w}_n^M$ from the computational model M .
 These are drawn from a representational space RS^M populated by the vectorial
 embeddings. At the same time, it is possible to derive a brain representation of the
 same words. This representation is the word-wise single trial activity recorded in the
 MEG dataset described in Section 4. Note that we did not average over time per trial,
 but that we kept the signal for each word as is. These activation samples also create a
 representational space RS^B populated by the same words $w_1, \dots, w_n$. At this point, the
 aim of the analysis is simply to measure how similar RS^M and RS^B are. In order to do
 so, we employ representational similarity analysis (RSA). Instead of directly mapping
 the two spaces, RSA compares pairwise similarities across different spaces.

We conducted RSA at the level of anatomical regions and using a sliding, partially
 overlapping, temporal window (width 30 ms, step 16 ms). In this way, we obtained a
 neural representation for each word, for each brain region and time window, which could
 be paired with a specific model-derived vectorial representation of the same word.

Given a model RS^M , we first obtained a representational similarity structure ss^M
 consisting of a $[n \times n]$ matrix where element ss_{ij}^M denotes the dissimilarity (computed as
 Euclidean distance) between embeddings $\vec{w}_i^M$ and $\vec{w}_j^M$ from RS^M of words w_i and w_j .
 Similarly it is possible to derive a series of representational structures ss^{Brt} for each
 anatomical region r and time window t . Element ss_{ij}^{Brt} quantifies the similarity between
 the brain activity in r and time t for word stimuli w_i and w_j . Fig 5 displays the
 representational similarity structure for brain activity in region r and time i (left), and
 for model M (right).

Fig 5. Similarity matrix for brain activity and for model. Similarity matrix for
 brain activity in region a and time i (left), and for model M (right). The matrices have
 been zoomed in to show their fine-grained structure and the labels of their rows and
 columns.

Fig 6. Representational Similarity Analysis (RSA). RSA is conducted by
 computing the similarity score between the brain space similarity matrices (left) and
 model similarity matrix (right). Here, the brain similarity matrix consists of the
 pairwise similarity scores among trials at anatomical region a and time window i .

Fig 7. RSA over time. RSA is conducted over time, meaning that a brain space
 similarity matrix is computed for each separate time window of the MEG trials.

The similarity score is estimated by taking Pearson’s correlation coefficient between the upper off-diagonal triangle of the $[n \times n]$ symmetric paired similarity matrices (ss^M and $ss^{B_{ai}}$) (Fig 6). These scores quantify the extent to which the similarity across stimuli is similarly represented by the model M and by brain activity in anatomical region a and time i . These measures are repeated across time t and anatomical regions r (Fig 7).

Therefore for each anatomical region, we obtained a representation of similarity between model and brain activity as a function of time from 0 to 500 ms after word onset (in windows of 30 ms, every 16 ms).

The analyses are conducted at the single-subject level up to this stage. The result of the temporal RSA for model M thus consists of 74 matrices, one for each subject, of size $[s \times u]$ where s is the number of anatomical regions and u the number of temporal windows.

Group-level analysis

Given 74 per-subject $[s \times u]$ matrices, we want to obtain a single group-level matrix describing the similarity between model M and brain activity across anatomical regions and over time.

Typically, group-level RSA results are obtained by averaging similarity matrices across subjects before computing the similarity score with the model similarity matrix. Cross-subject averaging requires that all matrices have the same size and that the row and columns of the matrices correspond to the same words across matrices. This is not the case for this dataset. First, the subjects are grouped in cohorts of about 17, each of which was presented with a different subset of the 360 stimulus sentences. This means that not all $ss^{B_{rt}}$ have rows and columns corresponding to the same words. Moreover, the trial selection procedure in the per-subject pre-processing consisted of discarding trials with irreparable artifacts. Therefore, even if two subjects were presented with the same set of sentences, there is a chance that they would have a different set of corresponding MEG trials due to the different occurrence of signal artifacts.

In our analysis, we computed t-statistics over subjects for each region/time combination independently. We computed a one-sided t-score for each of our

computational models ($W\vec{2}V$, $EL\vec{M}o$). We also computed a one-side t-score between the scores obtained by our aggregate models and word2vec ($EL\vec{M}o > W\vec{2}V$). The results of these t-statistics are thresholded at $p < 0.05$. Given the exploratory nature of this study, no statistic correction was applied in order not to obliterate the possible small effects detected by the RSA.

Results

The results of the analyses are split in two main parts. In the first part, we report the results of each of the two embedding models (word2vec and ELMo) separately. In the second part of the analyses, the similarity scores of word2vec and ELMo are contrasted to each other.

Results are provided at the whole-brain level, displaying the model–brain similarities at 5 distinct time points: 150, 250, 350, 450, and 550 milliseconds after word onset.

Models

In this section, we report the results of the RSA analyses of the two computational models separately.

The embedding model (Fig 8) which does not include contextual information, word2vec, returns lower similarities overall over time when correlated with brain activity, but more so from about 300 ms post word onset. For earlier latencies, word2vec shows significant activity in the left middle and inferior temporal gyri. Significant similarity with brain activity is also observed around 400 ms in the left posterior superior temporal gyrus.

Fig 8. Non-contextualized model (word2vec) This figure displays the results of the RSA between MEG activity and word2vec embeddings (t-values, $p < 0.05$).

The contextualized model (Fig 9), ELMo, instead exhibits an overall significant similarity with brain activity between 300 and 500 ms in the left frontal, prefrontal, and left anterior temporal regions. In particular, ELMo shows significant similarity with the left inferior temporal gyrus and the left anterior temporal cortex around 400 ms.

Fig 9. Contextualized model (ELMo) This figure displays the results of the RSA between MEG activity and ELMo embeddings (t-values, $p < 0.05$).

Comparison

In Fig 10 we show the results of a direct comparison between ELMo and word2vec ($ELMo > W2V$).

Fig 10. Comparison between contextualized (ELMo) and non-contextualized model (word2vec) This figure displays the comparison the correlations between MEG activity and ELMo and MEG activity and word2vec (t-values, $p < 0.05$).

ELMo shows significantly higher similarity to brain activity as compared to word2vec in the left anterior temporal lobe and in the left inferior frontal gyrus around 400 ms post word onset. The higher scores are also observed for the left middle frontal and left prefrontal regions around 500 ms.

Discussion

In Section 4, we observed that contextual and non-contextualized models yield qualitatively different results with regard to the timing and location of their similarity to MEG-recorded brain activity. In this section, we discuss the implications of these findings in light of the nature of the models and of the brain processing of natural language, as introduced in Sections 4 and 4.

We believe that computational word embedding models help in probing the nature of the neural representations correlated to memory retrieval and to integration. This is because they make the distinction between these two phenomena more computationally specific. When discussing the nature of retrieval and integration, Baggio & Hagoort (2011) [7] argue that approaches based on formal semantics might not be realistic models of how the brain implements these two operations. In agreement with Seuren (2009) [47], they state that formal semantics disregards natural language as a psychological phenomenon. They continue stating their desire to develop an account “that adheres to cognitive realism, in that it explains how language users derive meaning and how the human brain instantiates the neural architecture necessary to achieve this feat”. We believe that distributional semantic models, of which contextualized

embeddings are the most advanced version, have already proven their cognitive realism by being good models of human behavioral – e.g. semantic similarity judgment [48] – and neural [49, 50] data. Moreover, at the dawn of the field, distributional models – e.g., Latent Semantic Analysis [51] – were actually developed as cognitive models to answer questions on how children acquire word meaning, and how humans react to semantic similarity and relatedness. In light of the above considerations, we think that the models presented in the study might offer a cognitively realistic approximation of what is going on in the brain during memory retrieval and integration.

In the remainder of this section, we will discuss the effect of contextualization on the similarities between computational representations and brain activity (Section 4). We will specifically focus on the implications of these findings regarding the role of the anterior temporal lobe 4 and of activity peaking around 400 ms after stimulus onset (Section 4). We will also discuss the plausibility of the models chosen for the present study (Section 4).

The effect of contextualization on model-brain similarity

ELMo creates vectorial representation of words (embeddings) dependent on the context in which those words occur (Section 4). This contrasts with the nature of embeddings generated by a model such as word2vec (Section 4) that, once trained, always generates the same embeddings for a word regardless of its context.

Contextualized word embeddings generated by ELMo display higher similarity with brain activity in the frontal and anterior temporal regions of the brain around 300 and 400 ms post word onset; in other words, in areas and in a time frame compatible with integration processes. Integration refers to the integration of a linguistic unit (a word for instance) in the context provided by the other linguistic input in which it happens to be contained.

Non-contextualized word embeddings generated by word2vec exhibit a somewhat opposite behavior, showing higher similarity with earlier (100–200 ms) activity in lateral temporal regions. These regions are supposed to implement long-term memory (semantic memory) retrieval.

The fact that ELMo and word2vec embeddings display such different behaviors with

regard to brain activity can be reconciled with the division of labor predicated by
models such as the MUC model [28], especially concerning the distinction between
retrieval from memory and contextual integration.

The role of the anterior temporal lobe in integration

The fact that contextualized embeddings show higher similarity in the left anterior
temporal lobe might indicate that this region plays a role in integration. This is in line
with several experimental studies reporting the involvement of this area in lexical,
semantic and syntactic integration. Works such as Mazoyer & al. (1993) [29], Stowe &
al. (1998) [30], Friederici & al. (2000) [15], Humphries & al. (2006) [31], and Humphries
& al. (2007) [32] arrived at the conclusion that the anterior temporal lobe is somehow
involved in sentence-level integration, after observing an increase in activity in this area
during the presentation of sentences as compared to word lists. In addition, a series of
other studies have confirmed the role of ATP in processing composition by showing that
its activity is modulated by the type of syntactic relation holding between two words
being integrated [33–36, 36, 37].

The left ATP is also considered central in semantic memory. This was confirmed by
studies on patients suffering from semantic dementia [21, 22, 52, 53], and several
functional neuroimaging studies [23–26, 54]. These findings are summarized by
Patterson & al. 2007 [27], who hypothesized that concepts are represented by a network
of sensorimotor representations united by the left ATP.

Contextualized models do not show high correlations in the ATP at latencies that
are associated with memory-related processing: 100–200 ms after word presentation.
The results, therefore, seem to indicate that contextualized embeddings approximate
representations that have to do with the integration into context and not with lexical
retrieval from memory per se. If this were the case, we would have expected similar
correlations in the ATP both at memory retrieval (100–200 ms) and integration
latencies (400 ms). Conversely, a tentative interpretation with regard to the role of ATP
could be that activity in this region is different when semantic memory operations and
contextual integration processes are carried out.

The role of N400 in integration

The results presented in Section 4 do confirm not only the anatomical loci of memory and integration, but also provide indirect suggestions on the role of the N400 [41]. It has been debated in the literature whether the N400 is best characterized as playing a role in combinatorial (integration) or non-combinatorial (retrieval from memory) processes. Baggio & Hagoort (2011) [7] and Hagoort & al. (2008) [40] refrain from providing a stance on the matter because of the difficulty of devising convincing task-dependent experimental designs that are able to disentangle combinatorial and non-combinatorial semantic processes. Here again, we believe that the present computational modeling approach may provide an answer to the question. The results seem to point more towards a combinatorial process for the N400, given that the contextualized model maps to brain activity corresponding to its latencies. It is an indirect proof, but a proof pointing in this direction nonetheless. These results are in line with studies that link the role of the N400 to context processing [38,39].

The plausibility of bi-directional RNNs

ELMo is essentially a bi-directional recurrent neural network-based language model that integrates a word with its preceding and following context. Bi-directional recurrent language models seem to violate the assumption that human language processing proceeds left-to-right and word-by-word. Although this is trivially true for listening, it is worth noting that studies of reading behavior tend to describe a more nuanced situation. It is consistently reported in the eye-tracking of reading literature that between 15 and 20 % of eye movements proceed right-to-left (regressions) for left-to-right languages such as English or Dutch, and the opposite occurs for right-to-left languages such as Arabic [55]. Moreover, a number of “jump-ahead” eye movements are also commonly observed, indicating that humans either skip information that is deemed irrelevant for the processing of a linguistic item or that they look ahead in order to collect contextual information to the right of a word. This indicates that the preceding linguistic sequence is not always the only contextual material employed in the processing and interpretation of a word or of a sentence as a whole.

Conclusions

Recent developments in computational linguistics have created a new family of models which generate word embeddings that compute their representations with information derived from the context in which words are used. In this study, we have adopted one particular contextualized model, ELMo, as an approximation of the result of integration processes in the human brain during natural language comprehension. We contrasted ELMo with word2vec, a non-contextual embeddings model. Starting from the distinction between semantic memory retrieval (implemented in left temporal regions and activated around 200 ms after the onset of an incoming word) and word integration into context (carried out in left inferior frontal and left anterior temporal regions around 400 ms after word onset), we observed that non-contextualized models correlate with activity only in regions and latencies associated with semantic memory. In contrast, contextualized models correlate with activity in areas and latencies associated with word integration in context. These results confirm the functional and physiological distinction between memory and integration. Moreover, they provide some insight into the role of the left IFG, an area involved in integration and whose activity might temporarily store contextualized lexical representations. The results also point towards an involvement of the left anterior temporal lobe in integration, an area that was already linked to semantic combinatorial processes and which nonetheless received less attention in the theories of the architecture of the language system adopted in this study.

By highlighting a parallelism between models and brain activity, our results offer a contribution to the understanding of the division of labor at the cortical level between areas encoding lexical items in isolation and areas sensitive to the use of those items in context.

Bibliography

References

1. Damasio H, Grabowski TJ, Tranel D, Hichwa RD, Damasio AR. A neural basis for lexical retrieval. *Nature*. 1996;380:499–505.

2. Caramazza A. The brain's dictionary. *Nature*. 1996;380(485):1476 – 4687.
doi:<https://doi.org/10.1038/380485a0>.
3. Jackendoff R. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press; 2003. Available from:
<https://books.google.nl/books?id=228WWikKa-UC>.
4. Hagoort P, Levelt WJM. The Speaking Brain. *Science*. 2009;326(5951):372–373.
doi:10.1126/science.1181675.
5. Friederici AD. The Brain Basis of Language Processing: From Structure to Function. *Physiological Reviews*. 2011;91(4):1357–1392.
doi:10.1152/physrev.00006.2011.
6. Ullman MT. The biocognition of the mental lexicon. In: *The Oxford Handbook of Psycholinguistics*. Oxford University Press, Oxford; 2012.
7. Baggio G, Hagoort P. The balance between memory and unification in semantics: A dynamic account of the N400. *Language and Cognitive Processes*. 2011;26(9):1338–1367. doi:10.1080/01690965.2010.542671.
8. Mikolov T, Chen K, Corrado GS, Dean J. Efficient Estimation of Word Representations in Vector Space. *CoRR*. 2013;abs/1301.3781.
9. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. In: *Proc. of NAACL*; 2018.
10. Willems RM, Casasanto D. Flexibility in Embodied Language Understanding. In: *Front. Psychology*; 2011.
11. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*. 2018;.
12. Melamud O, Goldberger J, Dagan I. context2vec: Learning Generic Context Embedding with Bidirectional LSTM. In: *CoNLL*; 2016.
13. McCann B, Bradbury J, Xiong C, Socher R. Learned in Translation: Contextualized Word Vectors. In: *NIPS*; 2017.

14. Peters ME, Ammar W, Bhagavatula C, Power R. Semi-supervised sequence tagging with bidirectional language models. In: ACL; 2017.
15. Friederici AD, Meyer M, von Cramon DY. Auditory Language Comprehension: An Event-Related fMRI Study on the Processing of Syntactic and Lexical Information. *Brain and Language*. 2000;74:289–300.
16. Friederici AD. Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences*. 2002;6(2):78 – 84.
doi:[https://doi.org/10.1016/S1364-6613\(00\)01839-8](https://doi.org/10.1016/S1364-6613(00)01839-8).
17. Hagoort P. On Broca, brain, and binding: a new framework. *Trends in Cognitive Sciences*. 2005;9(9):416 – 423. doi:<http://dx.doi.org/10.1016/j.tics.2005.07.004>.
18. Schoffelen JM, Oostenveld R, Lam N, Udden J, Hultén A, Hagoort P. A 204-subject multimodal neuroimaging dataset to study language processing. *Scientific Data*. 2019;6. doi:10.1038/s41597-019-0020-y.
19. Binder JR, Desai RH, Graves WW, Conant LL. Where Is the Semantic System? A Critical Review and Meta-Analysis of 120 Functional Neuroimaging Studies. *Cerebral Cortex*. 2009;19(12):2767–2796. doi:10.1093/cercor/bhp055.
20. Binder JR, Desai RH. The neurobiology of semantic memory. *Trends in Cognitive Sciences*. 2011;15:527–536.
21. Hodges JR, Patterson K, Oxbury S, Funnell E. Semantic Dementia: progressive fluent aphasia with temporal lobe atrophy. *Brain*. 1992;115(6):1783–1806.
doi:10.1093/brain/115.6.1783.
22. Rogers TT, Ralph MAL, Garrard P, Bozeat S, McClelland JL, Hodges JR, et al. Structure and deterioration of semantic memory: a neuropsychological and computational investigation. *Psychological review*. 2004;111 1:205–35.
23. Tyler LK, Stamatakis EA, Bright P, Acres K, Abdallah S, Rodd JM, et al. Processing Objects at Different Levels of Specificity. *Journal of Cognitive Neuroscience*. 2004;16:351–362.

24. Bright P, Moss HE, Stamatakis EA, Tyler LK. The Anatomy of Object Processing: The Role of Anteromedial Temporal Cortex. *The Quarterly Journal of Experimental Psychology Section B*. 2005;58(3-4b):361–377. doi:10.1080/02724990544000013.
25. Moss HE, Rodd JM, Stamatakis EA, Bright P, Tyler LK. Anteromedial Temporal Cortex Supports Fine-grained Differentiation among Objects. *Cerebral Cortex*. 2004;15(5):616–627. doi:10.1093/cercor/bhh163.
26. Rogers TT, Hocking J, Noppeney U, Mechelli A, Gorno-Tempini ML, Patterson K, et al. Anterior temporal cortex and semantic memory: Reconciling findings from neuropsychology and functional imaging. *Cognitive, Affective, & Behavioral Neuroscience*. 2006;6(3):201–213. doi:10.3758/CABN.6.3.201.
27. Patterson K, Nestor PJ, Rogers TT. Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*. 2007;8:976–987.
28. Hagoort P. MUC (Memory, Unification, Control) and beyond. *Front Psychology*. 2013;4:416. doi:10.3389/fpsyg.2013.00416.
29. Mazoyer BM, Tzourio N, Frak V, Syrota A, Murayama N, Levrier O, et al. The Cortical Representation of Speech. *Journal of Cognitive Neuroscience*. 1993;5(4):467–479. doi:10.1162/jocn.1993.5.4.467.
30. Stowe LA, Broere CAJ, Paans AMJ, Wijers AA, Mulder G, Vaalburg W, et al. Localizing components of a complex task: sentence processing and working memory. *Neuroreport*. 1998;9 13:2995–2999.
31. Humphries C, Binder JR, Medler DA, Liebenthal E. Syntactic and Semantic Modulation of Neural Activity during Auditory Sentence Comprehension. *Journal of Cognitive Neuroscience*. 2006;18(4):665–679. doi:10.1162/jocn.2006.18.4.665.
32. Humphries C, Binder JR, Medler DA, Liebenthal E. Time course of semantic processes during sentence comprehension: An fMRI study. *NeuroImage*. 2007;36:924–932.

33. Baron SG, Thompson-Schill SL, Weber M, Osherson D. An early stage of conceptual combination: Superimposition of constituent concepts in left anterolateral temporal lobe. *Cognitive Neuroscience*. 2010;1(1):44–51. doi:10.1080/17588920903548751.
34. Baron SG, Osherson DN. Evidence for conceptual combination in the left anterior temporal lobe. *NeuroImage*. 2011;55:1847–1852.
35. Bemis DK, Pylkkänen L. Simple Composition: A Magnetoencephalography Investigation into the Comprehension of Minimal Linguistic Phrases. *Journal of Neuroscience*. 2011;31(8):2801–2814. doi:10.1523/JNEUROSCI.5003-10.2011.
36. Bemis DK, Pylkkänen L. Basic linguistic composition recruits the left anterior temporal lobe and left angular gyrus during both listening and reading. *Cerebral cortex*. 2013;23 8:1859–73.
37. Westerlund M, Kastner I, Kaabi MA, Pylkkänen L. The LATL as locus of composition: MEG evidence from English and Arabic. *Brain and Language*. 2015;141:124 – 134. doi:https://doi.org/10.1016/j.bandl.2014.12.003.
38. Berkum JJAV, Hagoort P, Brown CM. Semantic Integration in Sentences and Discourse: Evidence from the N400. *Journal of Cognitive Neuroscience*. 1999;11:657–671.
39. Hagoort P, van Berkum JJA. Beyond the sentence given. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2007;362 1481:801–811.
40. Hagoort P, Baggio G, Willems RM. Semantic Unification. In: Gazzaniga MS, editor. *The new cognitive neurosciences*, 4th edition. Cambridge (MA): MIT Press; 2009. p. 819–836.
41. Kutas M, Federmeier KD. Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual review of psychology*. 2011;62:621–647.

42. Hassan A, Mahmood A. Deep learning for sentence classification. 2017 IEEE Long Island Systems, Applications and Technology Conference (LISAT). 2017; p. 1–5.
43. Che W, Liu Y, Wang Y, Zheng B, Liu T. Towards Better UD Parsing: Deep Contextualized Word Embeddings, Ensemble, and Treebank Concatenation. In: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Brussels, Belgium: Association for Computational Linguistics; 2018. p. 55–64. Available from: <http://www.aclweb.org/anthology/K18-2005>.
44. Fares M, Kutuzov A, Oepen S, Velldal E. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In: Proceedings of the 21st Nordic Conference on Computational Linguistics. Gothenburg, Sweden: Association for Computational Linguistics; 2017. p. 271–276. Available from: <http://www.aclweb.org/anthology/W17-0237>.
45. Schoffelen JM, Hultén A, Lam N, Marquand AF, Uddén J, Hagoort P. Frequency-specific directed interactions in the human brain network for language. Proceedings of the National Academy of Sciences. 2017;doi:10.1073/pnas.1703155114.
46. Van Veen BD, Van Drongelen W, Yuchtman M, Suzuki A. Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. IEEE Transactions on biomedical engineering. 1997;44(9):867–880.
47. Seuren P. Language from within: Vol. 1. Language in cognition. Oxford, UK: Oxford University Press; 2009.
48. Baroni M, Lenci A. How we BLESSED distributional semantic evaluation. In: EMNLP 2011; 2011.
49. Mitchell TM, Shinkareva SV, Carlson A, min Chang K, Malave VL, Mason RA, et al. Predicting human brain activity associated with the meanings of nouns. Science. 2008;320 5880:1191–5.

50. Anderson AJ, Bruni E, Lopopolo A, Poesio M, Baroni M. Reading visually embodied meaning from the brain: Visually grounded computational models decode visual-object mental imagery induced by written text. *NeuroImage*. 2015;120:309–322.
51. Landauer TK, Dumais ST. A Solution to Plato’s Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*. 1997;104(2):211–240. doi:10.1037/0033-295X.104.2.211.
52. Hodges JR, Graham N, Patterson K. Charting the progression in semantic dementia: Implications for the organisation of semantic memory. *Memory*. 1995;3(3-4):463–495. doi:10.1080/09658219508253161.
53. Mummery CJ, Patterson K, Veltman DJ, Ashburner J, Frackowiak RS, Hodges JR. A voxel-based morphometry study of semantic dementia: relationship between temporal lobe atrophy and semantic memory. *Annals of neurology*. 2000;47 1:36–45.
54. Gauthier I, Anderson AW, Tarr MJ, Skudlarski P, Gore JC. Levels of categorization in visual recognition studied using functional magnetic resonance imaging. *Current Biology*. 1997;7(9):645 – 651. doi:[https://doi.org/10.1016/S0960-9822\(06\)00291-0](https://doi.org/10.1016/S0960-9822(06)00291-0).
55. Rayner K, Pollatsek A. *The Psychology of Reading*. Hillsdale: Lawrence Erlbaum Associates; 1995.

Figure 1

Figure 2

Figure 3

lexical retrieval

integration

Figure 4

Figure 5

Figure 6

Figure 7

Figure 8

Figure 9

Figure 10