Processing MWEs: Neurocognitive Bases of Verbal MWEs and Lexical Cohesiveness within MWEs

Shohini Bhattasali
Cornell University
Ithaca, NY, USA
sb2295@cornell.edu

Murielle Fabre
Cornell University / Ithaca, NY, USA
INSERM-CEA / Paris-Saclay, France
mf684@cornell.edu

John Hale
Cornell University / Ithaca, NY, USA
DeepMind / London, UK
jthale@cornell.edu

Abstract

Multiword expressions have posed a challenge in the past for computational linguistics since they comprise a heterogeneous family of word clusters and are difficult to detect in natural language data. In this paper, we present a fMRI study based on language comprehension to provide neuroimaging evidence for processing MWEs. We investigate whether different MWEs have distinct neural bases, e.g. if verbal MWEs involve separate brain areas from non-verbal MWEs and if MWEs with varying levels of cohesiveness activate dissociable brain regions. Our study contributes neuroimaging evidence illustrating that different MWEs elicit spatially distinct patterns of activation. We also adapt an association measure, usually used to detect MWEs, as a cognitively plausible metric for language processing.

1 Introduction

This study focuses on how Multiword Expressions are processed in the brain and provides a functional localization of different facets of MWEs using neuroimaging data. If MWEs are indeed non-compositional, then perhaps their comprehension proceeds through a single, unitary retrieval operation, rather than some kind of multistep compositional process. If we assume a single retrieval operation for these MWEs, how do the differences in their grammatical category affect their processing? Are they observable on the neuronal level?

Proceeding from this general hypothesis, this paper investigates the neural substrates of different types of MWEs and MWEs with different levels of compositionality. Firstly, verbal MWEs are distinguished from non-verbal MWEs and the neural bases of each are compared. Additionally, to model lexical cohesivenss of MWEs we use Pointwise Mututal Information, PMI (Church and Hanks, 1990), which is an association measure and traditionally used to identify MWEs. This gradient metric of cohesiveness within MWEs is correlated with brain activity to illustrate whether MWEs with varying degrees of compositionality evoke different patterns of activation in the brain. In this way, we provide further insight about MWE processing during natural language comprehension.

2 Background

2.1 Previous MWE Processing studies

MWE comprehension has been shown to be distinct from other kinds of language processing. For instance, it is well-established at the behavioural level that MWEs are produced and understood faster than matched control phrases due to their frequency, familiarity, and predictability (Siyanova-Chanturia and Martinez, 2014), in accordance with incremental processing (Hale, 2006). This would follow if MWEs were remembered as chunks, in the sense of Miller (1956) that was later formalised by Laird, Rosenbloom and Newell (1986; 1987).

Proceedings of the Joint Workshop on
Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), pages 6–17
Santa Fe, New Mexico, USA, August 25-26, 2018.

Eye-tracking and EEG work further documents this processing advantage across a wide range of MWE sub-types, e.g.

- Binomials (Siyanova-Chanturia et al., 2011b),
- Phrasal verbs (Yaneva et al., 2017),
- Complex prepositions (Molinaro et al., 2013; Molinaro et al., 2008),
- Nominal compounds (Molinaro and Carreiras, 2010; Molinaro et al., 2012),
- Lexical bundles (Tremblay and Baayen, 2010; Tremblay et al., 2011),
- Idioms (Underwood et al., 2004; Siyanova-Chanturia et al., 2011a; Strandburg et al., 1993; Laurent et al., 2006; Vespignani et al., 2010; Rommers et al., 2013).

For example, Siyanova-Chanturia et al. (2011b), found their eye-tracking results illustrate that binomial MWEs such as *bride and groom* are processed faster that the reversed three-word phrase *groom and bride*, due to the high-frequency nature of the former expression.

However, previous work has focused on a particular subtype of MWEs and to our knowledge, none of them have implemented a fMRI study of MWEs within a naturalistic text to either contrast between different categories of MWEs or model the cohesiveness within them. Recent computational work (Savary et al., 2017; Cholakov and Kordoni, 2016; Gharbieh et al., 2016; Uresova et al., 2016) has focused on verbal MWEs in order to identify them within a corpus, rather than study how they are processed in a naturalistic setting.

2.2 MWEs and Compositionality

The name MWE loosely groups a wide variety of linguistic phenomena including idioms, perfunctory greetings, character names, and personal titles. What unifies cases of MWEs is the absence of a wholly compositional linguistic analysis; they are "expressions for which the syntactic or semantic properties of the whole expression cannot be derived from its parts" (Sag et al., 2002). The naturalistic story used as a stimulus in this study includes various types of MWEs and some examples from the stimulus, *The Little Prince* are given below. The bold expressions were identified using a MWE analyzer, explained further in §4.2. Over half of the attestations in the text are headed by a verb and can be labelled as VPs (see §5.2.1). These encompass verb participle constructions, light verb constructions, and verb nominal constructions among others. The remaining attestations are a mixture of nominal compounds, greetings, personal titles, character names, and complex prepositions.

(1) So I thought **a lot** about the adventures of the jungle and **in turn**, I managed with a **coloured pencil** to make my first drawing.

(2) My **little fellow**, I don't know how to draw anything except **boa constrictors**, closed and open.

(3) When I drew the baobabs, I was spurred on by a **sense of urgency**.

(4) 'What are you doing there?', he said to the drinker who he found sitting **in silence** in front of **a number** of empty bottles and **a number** of full bottles.

(5) You must **see to it** that you regularly **pull out** the baobabs as soon as they can be told **apart from** the rose bushes to which they look very similar to when they are young.

(6) "**Good morning**", said the **little prince** politely, who then turned around, but saw nothing.

However, MWEs cannot be strictly binarized as compositional and non-compositional. These expressions fall along a graded spectrum of compositionality. To capture the varying degrees of compositionality within MWEs, we use an association measure, known as Pointwise Mutual Information (PMI). While PMI scores are commonly used in computational linguistics to identify MWEs as ngrams with higher scores are likely to be MWEs (Evert, 2008), in this study they are utilized as a gradient predictor to describe the the lexical cohesiveness of MWEs. Intuitively, its value is high when the word sequence under

consideration occurs more often together than one would have expected, based on the frequencies of the individual words (Manning et al., 1999). More formally, PMI is a log-ratio of observed and expected counts:

$$\text{PMI} = log_2 \frac{c(w_n^1)}{E(w_n^1)} \tag{1}$$

MWEs can receive positive or negative PMI scores which indicate cohesion or repulsion respectively between the words in a sequence (Church and Hanks, 1990). MWEs that receive a higher PMI score are seen as lexically more cohesive, which is interpreted as more noncompositional (less compositional). Thus, these scores are repurposed in this study to describe the cohesive and noncompositional aspect of MWEs and utilized to obtain a quantifiable metric to correlate with the fMRI signal. Krenn (2000) also suggests that association measures such as PMI and Dice's coeeficient (Dice, 1945; Sørensen, 1948; Smadja et al., 1996) are better-suited to identify high-frequency collocations whereas other association measures such as log-likelihood are better at detecting medium to low frequency collocations. Since MWEs are inherently high-frequency collocations, we chose PMI as a metric to describe the strength of association between these word clusters.

3 Research Questions

To summarize, this study investigates the following:

- Are the differences between the grammatical categories of MWEs observable at the cerebral level? Does processing of verbal MWEs implicate separate brain areas from non-verbal MWEs? Specifically, if the strong relationship between verbs and their arguments are encoded in different brain areas compared to non-verbal MWEs featuring no argumental structure? (c.f. Analysis 1 in §5.2.1)

- Do MWEs with varying levels of cohesiveness tap into different cognitive resources? For example, are MWEs with higher PMI scores processed differently from MWEs with lower scores? Do they activate dissociable brain regions? (c.f. Analysis 2 in §5.2.2)

4 fMRI study

4.1 Method

We follow Brennan et al., (2012) in using a spoken narrative as a stimulus. Participants hear the story over headphones while they are in the scanner. The sequence of neuroimages collected during their session becomes the dependent variable in a regression against word-by-word predictors, derived from the text of the story.

4.2 Stimuli & MWE Identification

The audio stimulus was Antoine de Saint-Exupéry's *The Little Prince*, translated by David Wilkinson and read by Nadine Eckert-Boulet. It constitutes a fairly lengthy exposure to naturalistic language, comprising 15,388 words and lasting over an hour and a half.

Within this text, 742 MWEs were identified using a transition-based MWE analyzer (Al Saied et al., 2017). Al Saied et al. use unigram and bigram features, word forms, POS tags and lemmas, in addition to features such as transition history and report an average F-score 0.524 for this analyzer across 18 different languages which reflects robust cross-linguistic performance. For an illustrated example of the MWE identification process with this analyzer, please see the Appendix. The analyzer was trained on examples from the Children's Book Test, CBT (Hill et al., 2015) from the Facebook bAbI project to keep the genre consistent with our literary stimulus. This corpus consists of text passages that are drawn from the Children's section of Project Gutenberg, a free online text repository. External lexicons were used to supplement the MWEs found with the analyzer. The external lexicons included the Unitex lexicon (Paumier et al., 2009), the SAID corpus (Kuiper et al., 2003), the Cambridge International Dictionary of Idioms (White, 1998), and the Dictionary of American Idioms (Makkai et al., 1995). While 742 MWEs

might seem like a restricted sample, this data was acquired with experimental constraints since our fMRI study was almost two hours long which is on the longer end for similar neuroimaging studies.

4.3 Participants

Participants were fifty-one volunteers (32 women and 19 men, 18-37 years old) with no history of psychiatric, neurological, or other medical illness or history of drug or alcohol abuse that might compromise cognitive functions. All strictly qualified as right-handed on the Edinburgh handedness inventory (Oldfield, 1971). They self-identified as native English speakers and gave their written informed consent prior to participation, in accordance with Cornell University IRB guidelines.

4.4 Presentation

After giving their informed consent, participants were familiarized with the MRI facility and assumed a supine position on the scanner gurney. The presentation script was written in PsychoPy peirce:2007. Auditory stimuli were delivered through MRI-safe, high-fidelity headphones (Confon HP-VS01, MR Confon, Magdeburg, Germany) inside the head coil. The headphones were secured against the plastic frame of the coil using foam blocks. Using a spoken recitation of the US Constitution, an experimenter increased the volume until participants reported that they could hear clearly. Participants then listened passively to the audio storybook for 1 hour 38 minutes. The story had nine chapters and at the end of each chapter the participants were presented with a multiple-choice questionnaire with four questions (36 questions in total), concerning events and situations described in the story. These questions served to confirm participants' comprehension. They were viewed via a mirror attached to the head coil and answered through a button box. The entire session lasted around 2.5 hours.

4.5 Data Collection

Imaging was performed using a 3T MRI scanner (Discovery MR750, GE Healthcare, Milwaukee, WI) with a 32-channel head coil at the Cornell MRI Facility. Blood Oxygen Level Dependent (BOLD) signals were collected using a T2-weighted echo planar imaging (EPI) sequence (repetition time: 2000 ms, echo time: 27 ms, flip angle: 77deg, image acceleration: 2X, field of view: 216 x 216 mm, matrix size 72 x 72, and 44 oblique slices, yielding 3 mm isotropic voxels). Anatomical images were collected with a high resolution T1-weighted (1 x 1 x 1 mm^3 voxel) with a Magnetization-Prepared RApid Gradient-Echo (MP-RAGE) pulse sequence.

5 Data Analysis

5.1 Preprocessing

fMRI data is acquired with physical, biological constraints and preprocessing allows us to make adjustments to improve the signal to noise ratio. Primary preprocessing steps were carried out in AFNI version 16 (Cox, 1996) and include motion correction, coregistration, and normalization to standard MNI space. After the previous steps were completed, ME-ICA (Kundu et al., 2012) was used to further preprocess the data. ME-ICA is a denoising method which uses Independent Components Analysis to split the T2*-signal into BOLD and non-BOLD components. Removing the non-BOLD components mitigates noise due to motion, physiology, and scanner artifacts (Kundu et al., 2017).

5.2 Statistical Analysis

The General Linear Model (GLM) typically used in fMRI data analysis is a hierarchical model with two levels (Poldrack et al., 2011). At the first level, the data for each subject is modelled separately to calculate subject-specific parameter estimates and within-subject variance such that for each subject, a regression model is estimated for each voxel against the time series. The second-level model takes subject-specific parameter estimates as input. It uses the between-subject variance to make statistical inferences about the larger population.

The research questions presented above in §3 motivate two statistical analyses. The first analysis localizes verbal MWEs and non-verbal MWEs to see if they activate spatially different networks in the

brain. The second analysis investigates MWEs along a quantitative gradient of lexical cohesion. Both analyses employ the GLM, and were carried out using SPM12 (Friston et al., 2007). The predictors were convolved using the canonical HRF in SPM. For both of these analyzes, the MWE candidates were taken to be the expressions from the transition-based analyzer (as described in §4.2).

5.2.1 Analysis 1: Verbal MWEs vs. Non-verbal MWEs

We regressed the word-by-word predictors described below against fMRI timecourses recorded during passive story-listening in a whole-brain analysis. For each of the 15,388 words in the story, their timestamps were estimated using Praat TextGrids (Boersma, 2002). MWEs were identified, as described in §4.2 and the presence/absence of verbal expression yielded two categories of MWEs (i.e. 56% verbal vs. 44% non-verbal). The Stanford POS tagger and the NLTK POS tagger were used to annotate the words within the MWEs with their grammatical categories (Bird and Loper, 2004; Manning et al., 2014). Additionally, we entered four regressors of non-interest into the GLM analysis (SPM12): word-offset, word frequency, pitch, intensity which serve to improve the sensitivity, specificity and validity of activation maps (Bullmore et al., 1999; Lund et al., 2006). To control for sentence-level and phrase-level compositional processes, we included a regressor formalizing syntactic structure building based on a bottom-up parsing algorithm (Hale, 2014), as determined by the Stanford parser (Klein and Manning, 2003). Controlling for structural composition allows us to isolate and focus our investigation on noncompositional processing, as in MWEs. These regressors were not orthogonalized.

5.2.2 Analysis 2: Cohesiveness within MWEs

Analysis 2 uses the same predictors as in Analysis 1, except that the categorical indicators for MWEs is replaced with the gradient predictor, PMI. All the 742 MWEs that were annotated with a 1 in Analysis 1 are in Analysis 2 marked with their PMI score. This score is based on corpus frequency counts from the Corpus of Contemporary English (Davies, 2008), and were calculated using `mwetoolkit` (Ramisch et al., 2010; Ramisch, 2012). These regressors were also not orthogonalized.

5.2.3 Group-level Analysis

In the second-level group analysis, each contrast was analyzed separately at the group-level. An 8 mm FWHM Gaussian smoothing kernel was applied on the contrast images from the first-level analysis to counteract inter-subject anatomical variation. All the group-level results reported in the next section underwent FWE voxel correction for multiple comparisons which resulted in T-scores > 5.3.

6 Results

Behavioural results of the comprehension task showed attentive listening to the auditory story presentation. Across 51 participants, average accurate responses to the comprehension questions was 90% (SD = 3.7%).

6.1 Group-level results for Verbal MWEs vs Non-verbal MWEs

The main effect for presence of MWEs elicited activation mainly in bilateral Supramarginal Gyrus, right Angular Gyrus, right MFG, and right Precuneus Cortex (Fig. 1A). Whole-brain contrasts show that these two types of MWEs activate different brain regions with no overlap. Verbal MWEs appear right-lateralized compared to non-verbal ones in IPL and in IFG triangularis (Fig.1B). The opposite contrast yielded a mostly right-lateralized and wider pattern of activation, including bilateral Supramarginal Gyrus extending to STG and right SMA together with smaller activation clusters in Pars Opercularis and MTG (Fig. 1B). Contrasts were inclusively masked with the main effect of all MWEs.

6.2 Group-level results for Lexical Cohension with MWEs

Increasing cohesiveness, as seen through positive activation with PMI (Fig. 2, in purple), elicits the Precuneus and Supplementary Motor Area, while decreasing cohesiveness, as seen through negative activation with PMI (Fig. 2, in orange), correlates with activity in well-known nodes of the language network, such as Broca's area and the posterior Temporal Gyrus.

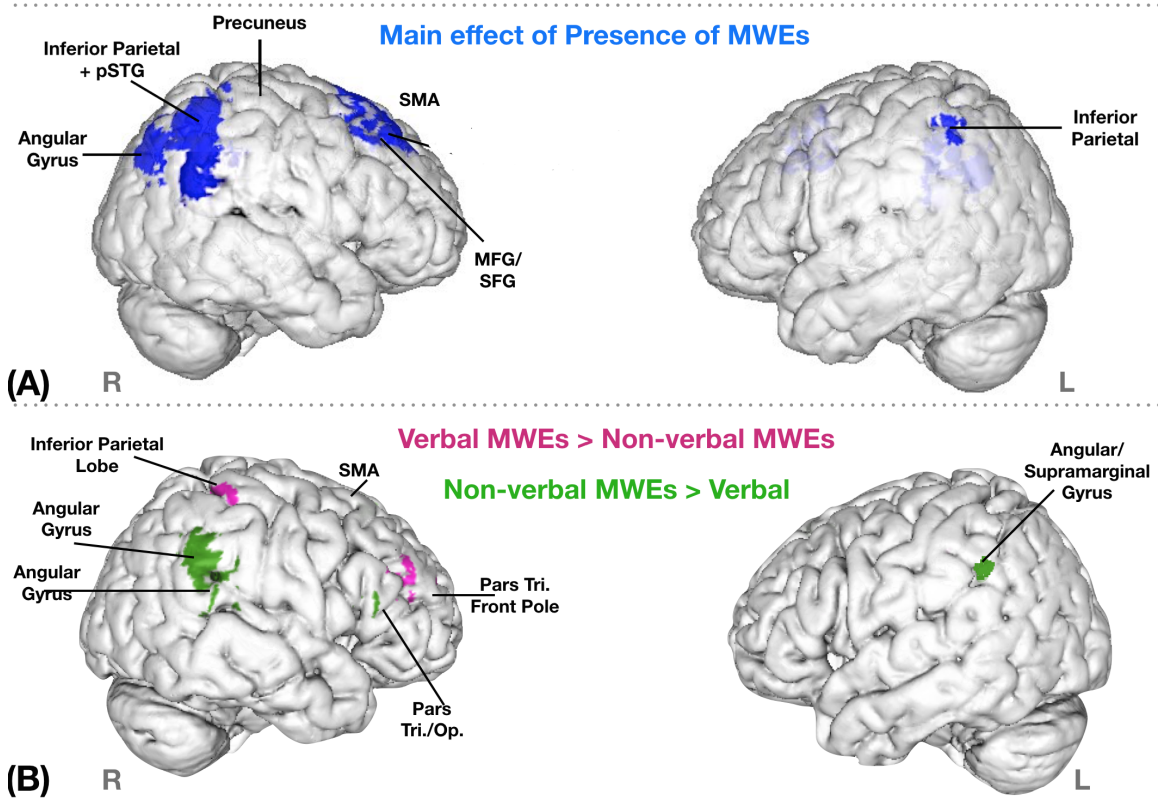

Figure 1: (A): Whole brain main effect for MWEs in blue. (B): Contrast images with significant clusters for [Verbal MWEs > Non-verbal MWEs] in pink and for [Nonverbal MWEs > Verbal MWEs] in green.

Regions	Cluster size (in voxels)	MNI Coordinates x	y	z	p-value (corrected)	T-score (peak level)
Verbal MWEs >Non-verbal MWEs						
R IFG Pars Triangularis	71	46	36	14	0.000	7.38
R Inferior Parietal Lobule	57	50	-40	52	0.002	6.38
Non-verbal MWEs >Verbal MWEs						
R Angular Gyrus	585	56	-42	14	0.000	9.43
R Supplementary Motor Area	235	12	20	60	0.000	8.91
L Cerebellum	58	-22	-72	-30	0.002	7.85
L Supramarginal Gyrus	32	-60	-50	34	0.001	6.50
R IFG Pars Triangularis/Opercularis	28	56	22	8	0.001	6.51

Table 1: Significant cluster for contrasts between verbal MWEs and non-verbal MWEs after FWE voxel correction for multiple comparisons with $p < 0.05$. Peak activation is given in MNI Coordinates.

7 Discussion & Further Work

The results from Analysis 1 provide evidence that MWEs activate areas consistently reported as the lexical semantic network, such as Supramarginal, Parietal areas, and SMA (Binder et al., 2009). MWEs mostly implicate a right-lateralized network while contrastively, compositional processes have been essentially linked to left lateralization (Friederici and Gierhan, 2013; Bemis and Pylkkänen, 2013; Bemis and Pylkkänen, 2011). Previous findings also show that the bilateral Supramarginal Gyrus is sensitive to co-occurence frequency of word combinations as reported previously for semantically meaningful and frequent word-pairs (Graves et al., 2010; Price et al., 2015).

Additionally, the significant clusters for verbal and non-verbal MWEs illustrate spatially distinct pat-

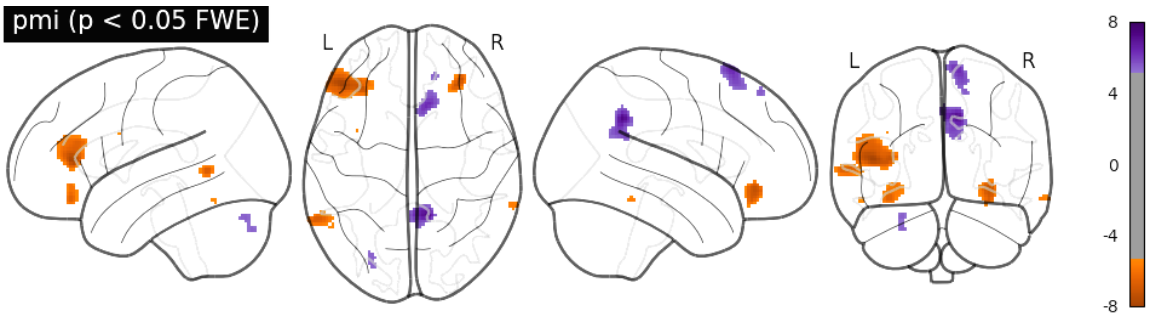

Figure 2: Significant cluster for the increasing and decreasing cohesion measure of MWEs after FWE voxel correction for multiple comparisons with $p < 0.05$ and cluster-extent threshold ($k > 50$) for display purposes. Peak activation is given in MNI Coordinates.

Regions for PMI	Cluster size (in voxels)	MNI Coordinates x	y	z	p-value (corrected)	T-score (peak-level)
Correlated with increasing MWE cohesion						
R Precuneus Cortex	263	10	-50	36	0.000	7.30
R Precuneus Cortex		10	-48	24	0.000	5.84
R Superior Frontal Gyrus/Supplementary Motor Area (BA6)	154	10	22	68	0.000	6.34
Correlated with decreasing MWE cohesion						
L IFG Pars Triangularis	448	-46	36	8	0.000	8.01
R IFG Pars Orbitalis/Middle Frontal Gyrus	117	32	38	-12	0.000	7.29
L Posterior Middle Temporal Gyrus	53	-62	-52	2	0.000	6.76
L IFG Pars Orbitalis/Middle Frontal Gyrus	67	-36	38	-16	0.000	6.62

Table 2: Significant cluster for the increasing and decreasing cohesion measure of MWEs after FWE voxel correction for multiple comparisons with $p < 0.05$. Peak activation is given in MNI Coordinates.

terns of activation and a dorso-ventral gradient is observed in Brocas area for verbal versus non-verbal MWEs. Activation patterns for verbal MWEs suggest that verb-argument selectional relations in frequent verbal expressions exclusively involve right hemisphere activity in Brocas area and IPL.

In the case of non-verbal MWEs, we do not make a strong conclusion since it is a mixed bag of nominal compounds, complex prepositions, greetings, personal titles among other types. We did not contrast between verbal and nominal MWEs since our dataset is skewed towards verbal MWEs and we have very few attestations of nominal MWEs in the text ($< 7\%$).

Our results from Analysis 2 show that highly cohesive MWEs implicate the Precuneus and the SMA, suggesting that only truly lexicalized linguistic expressions rely on these areas rather than traditional frontal and temporal nodes of the language network. These areas have been implicated in memory and naming tasks (Crosson, 2013; Halsband et al., 2002). Less cohesive MWEs activate core areas of the language network implicated in composition (Fedorenko et al., 2016; Friederici and Gierhan, 2013; Pallier et al., 2011; Snijders et al., 2009) which suggests that less cohesive MWEs are processed compositionally and are not retrieved as a unit.

Apart from an association measure like PMI, there are alternate approaches to describes MWEs such as word space models (based on distributional semantics) which could also serve as a metric of compositionality for these noncompositional word clusters. This type of metric would utilize the distributional patterns of words collected over large text data to represent semantic similarity between words in terms of spatial proximity (Sahlgren, 2006). However, in the current study we were not trying to model the semantic opacity of these expressions but that could be an area to explore in the future to investigate another aspect of MWEs.

This study only included native speakers of English as participants and is part of a larger project investigating MWEs cross-linguistically to compare if they are processed similarly. Another future research

direction would be to replicate the same experiment with non-native speakers to study how early or late acquisition of English would impact the neural bases recruited in processing these noncompositional expressions.

Another approach to illustrate this gamut of compositionality would be to compare a compositional expression like a VP against a noncompositional verbal MWE (e.g. *kick the ball* vs. *kick the bucket*. Morphosyntactically, these would be structurally similar yet they should be processed differently if our hypothesis about the neurocognitive mechanisms underlying language processing is correct. Based on our prediction, the neuroimaging data should illustrate a spatial dissociation between compositional VPs and noncompositional verbal MWEs.

8 Conclusion

Our results point to a spatial differentiation between verbal MWEs and non-verbal MWEs since they localize to different areas of the brain. Thus, this study provides neuroimaging evidence of different types of MWEs. Additionally, it also illustrates that the grammatical category of the words inside MWEs is crucial to how they are processed in the brain. For example, in the verbal MWEs scenario, the word clusters headed by a verb activate spatially different regions from non-verbal MWEs, plausibly due to the inherent argument structure present in verbal MWEs. Furthermore, this result illustrates that even within these noncompositional verbal expression, there is an aspect of argument structure composition within its subparts.

Furthermore using PMI as a gradient predictor shows that highly cohesive MWEs and less cohesive MWEs tap into different cognitive resources, as evidenced through their separate neural correlates. This suggests a difference between processing truly lexicalized MWEs in contrast to MWEs which are possibly analyzed compositionally. Lastly, one of the main contributions of this study is in repurposing PMI, an association measure to describe MWEs in terms of cohesion and thus showing that they are a cognitively informative metric to model cohesiveness and compositionality within word clusters in natural language.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1607441.

References

Hazem Al Saied, Marie Candito, and Matthieu Constant. 2017. The ATILF-LLF system for the PARSEME Shared Task: a Transition-based Verbal Multiword Expression Tagger. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 127–132, Valencia, Spain, April. Association for Computational Linguistics.

Douglas K Bemis and Liina Pylkkänen. 2011. Simple composition: A magnetoencephalography investigation into the comprehension of minimal linguistic phrases. *The Journal of Neuroscience*, 31(8):2801–2814.

Douglas K Bemis and Liina Pylkkänen. 2013. Basic linguistic composition recruits the left anterior temporal lobe and left angular gyrus during both listening and reading. *Cerebral Cortex*, 23(8):1859–1873.

Jeffrey R Binder, Rutvik H Desai, William W Graves, and Lisa L Conant. 2009. Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, 19(12):2767–2796.

Steven Bird and Edward Loper. 2004. Nltk: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 31. Association for Computational Linguistics.

Paul Boersma. 2002. *Praat, a system for doing phonetics by computer*. Glot International.

ET Bullmore, MJ Brammer, S Rabe-Hesketh, VA Curtis, RG Morris, SCR Williams, T Sharma, and PK McGuire. 1999. Methods for diagnosis and treatment of stimulus-correlated motion in generic brain activation studies using fmri. *Human brain mapping*, 7(1):38–48.

Kostadin Cholakov and Valia Kordoni. 2016. Using word embeddings for improving statistical machine translation of phrasal verbs. In *Proceedings of the 12th Workshop on Multiword Expressions*, pages 56–60.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Robert W. Cox. 1996. Afni: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical research*, 29(3):162–173.

Bruce Crosson. 2013. Thalamic mechanisms in language: A reconsideration based on recent findings and concepts. *Brain and Language*, 126(1):73–88.

Mark Davies. 2008. *The Corpus of Contemporary American English (COCA): 560 million words, 1990–present*. BYE, Brigham Young University.

Lee R Dice. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Stefan Evert. 2008. Corpora and collocations. In Anke Lüdeling and Merja. Kytö, editors, *Corpus linguistics: an international handbook*, pages 1212–1248. W. de Gruyter, Berlin. article number 58.

Evelina Fedorenko, Terri L Scott, Peter Brunner, William G Coon, Brianna Pritchett, Gerwin Schalk, and Nancy Kanwisher. 2016. Neural correlate of the construction of sentence meaning. *Proceedings of the National Academy of Sciences*, 113(41):E6256–E6262.

Angela D Friederici and Sarah ME Gierhan. 2013. The language network. *Current Opinion in Neurobiology*, 23(2):250–254.

K.J. Friston, J. Ashburner, S.J. Kiebel, T.E. Nichols, and W.D. Penny, editors. 2007. *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Academic Press.

Waseem Gharbieh, Virendra Bhavsar, and Paul Cook. 2016. A word embedding approach to identifying verb-noun idiomatic combinations. In *Proceedings of the 12th Workshop on Multiword Expressions*, pages 112–118.

William W Graves, Jeffrey R Binder, Rutvik H Desai, Lisa L Conant, and Mark S Seidenberg. 2010. Neural correlates of implicit and explicit combinatorial semantic processing. *Neuroimage*, 53(2):638–646.

John Hale. 2006. Uncertainty about the rest of the sentence. *Cognitive Science*, 30(4):643–672.

John T Hale. 2014. *Automaton theories of human sentence comprehension*. CSLI Publications.

U Halsband, BJ Krause, H Sipilä, M Teräs, and A Laihinen. 2002. Pet studies on the memory processing of word pairs in bilingual finnish–english subjects. *Behavioural brain research*, 132(1):47–57.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The Goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*.

Dan Klein and Christopher D Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

Brigitte Krenn. 2000. Empirical implications on lexical association measures. In *Proceedings of The Ninth EURALEX International Congress*.

Koenraad Kuiper, Heather McCann, Heidi Quinn, Therese Aitchison, and Kees van der Veer. 2003. Syntactically Annotated Idiom Dataset (SAID) LDC2003T10. In *Linguistic Data Consortium*, Philadelphia.

Prantik Kundu, Souheil J Inati, Jennifer W Evans, Wen-Ming Luh, and Peter A Bandettini. 2012. Differentiating bold and non-bold signals in fmri time series using multi-echo epi. *Neuroimage*, 60(3):1759–1770.

Prantik Kundu, Valerie Voon, Priti Balchandani, Michael V. Lombardo, Benedikt A. Poser, and Peter A. Bandettini. 2017. Multi-echo fmri: A review of applications in fmri denoising and analysis of bold signals. *NeuroImage*, 154:59 – 80. Cleaning up the fMRI time series: Mitigating noise with advanced acquisition and correction strategies.

John Laird, Paul Rosenbloom, and Allen Newell. 1986. Chunking in Soar, anatomy of a general learning mechanism. *Machine Learning*, 1.

Jean-Paul Laurent, Guy Denhières, Christine Passerieux, Galina Iakimova, and Marie-Christine Hardy-Baylé. 2006. On understanding idiomatic language: The salience hypothesis assessed by ERPs. *Brain Research*, 1068(1):151–160.

Torben E Lund, Kristoffer H Madsen, Karam Sidaros, Wen-Lin Luo, and Thomas E Nichols. 2006. Non-white noise in fmri: does modelling have an impact? *Neuroimage*, 29(1):54–66.

Adam Makkai, M. T. Boatner, and J. E. Gates. 1995. *A Dictionary of American idioms*. ERIC.

Christopher D Manning, Hinrich Schütze, et al. 1999. *Foundations of statistical natural language processing*, volume 999. MIT Press.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

George A. Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2):81–97.

Nicola Molinaro and Manuel Carreiras. 2010. Electrophysiological evidence of interaction between contextual expectation and semantic integration during the processing of collocations. *Biological Psychology*, 83(3):176–190.

Nicola Molinaro, Francesco Vespignani, Paolo Canal, Sergio Fonda, and Cristina Cacciari. 2008. Cloze probability does not only affect N400 amplitude: The case of complex prepositions. *Psychophysiology*, 45(6):1008–1012.

Nicola Molinaro, Manuel Carreiras, and Jon Andoni Duñabeitia. 2012. Semantic combinatorial processing of non-anomalous expressions. *Neuroimage*, 59(4):3488–3501.

Nicola Molinaro, Paolo Canal, Francesco Vespignani, Francesca Pesciarelli, and Cristina Cacciari. 2013. Are complex function words processed as semantically empty strings? A reading time and ERP study of collocational complex prepositions. *Language and Cognitive Processes*, 28(6):762–788.

Richard C Oldfield. 1971. The assessment and analysis of handedness: the edinburgh inventory. *Neuropsychologia*, 9(1):97–113.

Christophe Pallier, Anne-Dominique Devauchelle, and Stanislas Dehaene. 2011. Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, 108(6):2522–2527.

Sébastien Paumier, Takuya Nakamura, and Stavroula Voyatzi. 2009. Unitex, a corpus processing system with multi-lingual linguistic resources. *eLEX2009*, page 173.

Russell A Poldrack, Jeanette A Mumford, and Thomas E Nichols. 2011. *Handbook of functional MRI data analysis*. Cambridge University Press.

Amy R Price, Michael F Bonner, Jonathan E Peelle, and Murray Grossman. 2015. Converging evidence for the neuroanatomic basis of combinatorial semantics in the angular gyrus. *Journal of Neuroscience*, 35(7):3276–3284.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010. mwetoolkit: a Framework for Multiword Expression Identification. In *LREC*, volume 10, pages 662–669.

Carlos Ramisch. 2012. A generic framework for multiword expressions treatment: From acquisition to applications. In *Proceedings of ACL 2012 Student Research Workshop*, pages 61–66. Association for Computational Linguistics.

Joost Rommers, Ton Dijkstra, and Marcel Bastiaansen. 2013. Context-dependent Semantic Processing in the Human Brain: Evidence from Idiom Comprehension. *Journal of Cognitive Neuroscience*, 25(5):762–776.

Paul S. Rosenbloom and Allen Newell. 1987. Learning by chunking: A production-system model of practice. In *Production System Models of Learning and Development*, pages 221–286. MIT Press.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15. Springer.

Magnus Sahlgren. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, and Ivelina Stoyanova. 2017. The parseme shared task on automatic identification of verbal multiword expressions.

Anna Siyanova-Chanturia and Ron Martinez. 2014. The idiom principle revisited. *Applied Linguistics*, 36(5):549–569.

Anna Siyanova-Chanturia, Kathy Conklin, and Norbert Schmitt. 2011a. Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research*, 27(2):251–272.

Anna Siyanova-Chanturia, Kathy Conklin, and Walter JB Van Heuven. 2011b. Seeing a phrase time and again matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(3):776.

Frank Smadja, Kathleen R McKeown, and Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational linguistics*, 22(1):1–38.

Tineke M Snijders, Theo Vosse, Gerard Kempen, Jos JA Van Berkum, Karl Magnus Petersson, and Peter Hagoort. 2009. Retrieval and unification of syntactic structure in sentence comprehension: an fmri study using word-category ambiguity. *Cerebral cortex*, 19(7):1493–1503.

Thorvald Sørensen. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. *Biol. Skr.*, 5:1–34.

Robert J Strandburg, James T Marsh, Warren S Brown, Robert F Asarnow, Donald Guthrie, and Jerilyn Higa. 1993. Event-related potentials in high-functioning adult autistics: Linguistic and nonlinguistic visual information processing tasks. *Neuropsychologia*, 31(5):413–434.

Antoine Tremblay and R Harald Baayen. 2010. Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. *Perspectives on formulaic language: Acquisition and communication*, pages 151–173.

Antoine Tremblay, Bruce Derwing, Gary Libben, and Chris Westbury. 2011. Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning*, 61(2):569–613.

Geoffrey Underwood, Norbert Schmitt, and Adam Galpin. 2004. The eyes have it. *Formulaic Sequences: Acquisition, Processing, and Use*, 9:153.

Zdenka Uresova, Eduard Bejček, and Jan Hajic. 2016. Inherently pronominal verbs in czech: Description and conversion based on treebank annotation. In *Proceedings of the 12th Workshop on Multiword Expressions*, pages 78–83.

Francesco Vespignani, Paolo Canal, Nicola Molinaro, Sergio Fonda, and Cristina Cacciari. 2010. Predictive Mechanisms in Idiom Comprehension. *Journal of Cognitive Neuroscience*, 22(8):1682–1700.

Victoria Yaneva, Shiva Taslimipoor, Omid Rohanian, et al. 2017. Cognitive processing of multiword expressions in native and non-native speakers of English: Evidence from gaze data. In *International Conference on Computational and Corpus-Based Phraseology*, pages 363–379. Springer.

Appendix

Overview of the MWE identification, as per Al Saied et al., (2017):

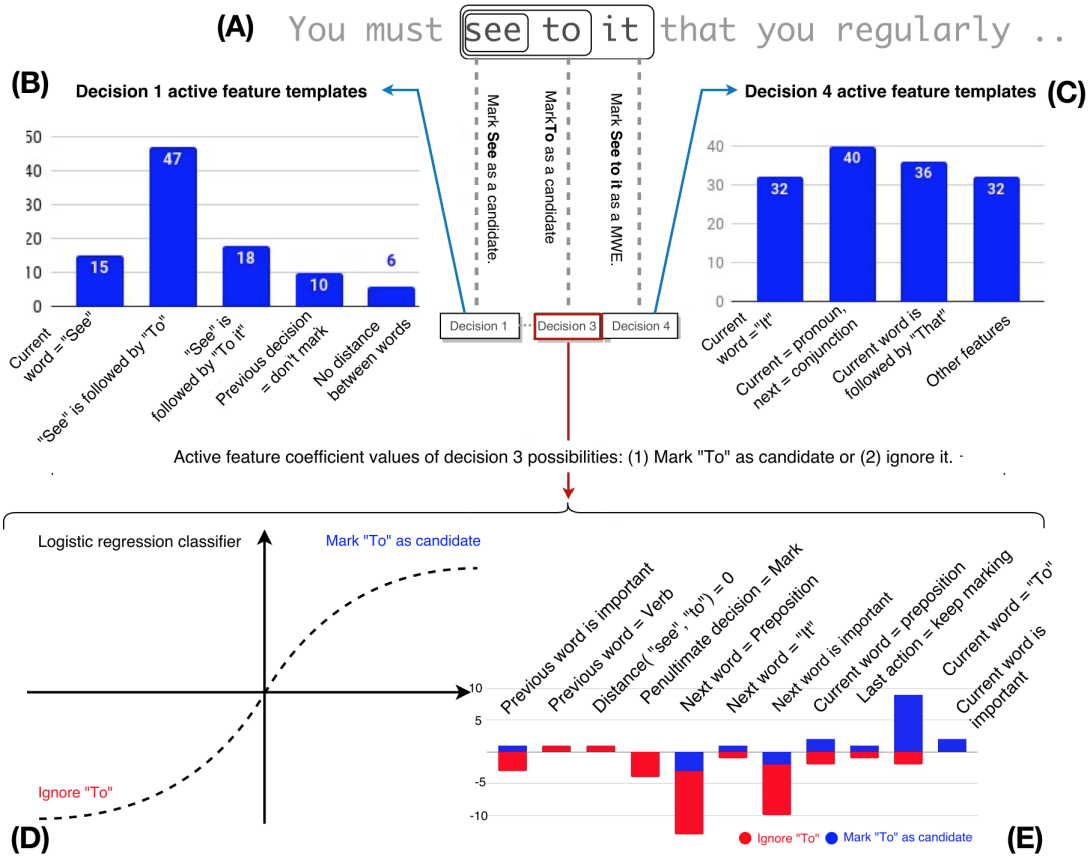

Figure 1: Identifying the multiword expression *see to it*. Panel (A) shows the context in which the MWE occurs. Identification in this case proceeds on the basis of four Decisions, numbered 1 through 4. The first three Decisions mark see , to and it respectively as candidate words. With the fourth Decision, the entire MWE is identified. The sorts of text-properties influencing Decisions 1 and 4 are shown in panels (B) and (C) respectively. These feature templates encourage the probabilistic classifier (panel D) to either mark or not. Panel (E) offers a closer look at the word to in terms of particular features that either encourage or discourage marking. Because the coefficient values are higher on the features that favour marking, the classifier chooses to mark to as a candidate for inclusion in the MWE.