Modeling Conventionalization and Predictability within MWEs at the Brain Level

Shohini Bhattasali *
Dept. of Linguistics – UMIACS
University of Maryland
shohini@umd.edu

Murielle Popa–Fabre *
INRIA – University of Paris
ALMANACH – LLF
murielle.fabre@inria.fr

Christophe Pallier
Cognitive Neuroimaging Lab
INSERM-CEA
christophe@pallier.org

John Hale
Dept. of Linguistics
University of Georgia
jthale@uga.edu

Abstract

While expressions have traditionally been binarized as compositional and noncompositional in linguistic theory, Multiword Expressions (MWEs) demonstrate finer-grained distinctions. Using Association Measures like Pointwise Mutual Information and Dice's Coefficient, MWEs can be characterized as having different degrees of conventionalization and predictability. Our goal is to investigate how these gradiences could reflect cognitive processes. In this study, fMRI recordings of naturalistic narrative comprehension is used to probe to what extent these computational measures and the cognitive processes they could operationalize are observable during on-line sentence processing. Our results show that Dice's Coefficent, representing lexical predictability, is a better predictor of neural activation for processing MWEs. Overall our experimental approach demonstrates how we can test the cognitive plausibility of computational metrics by comparing it against neuroimaging data.

1 Introduction

Multiword Expressions (MWEs) are word clusters or expressions formed by more than a single word. Siyanova-Chanturia (2013) provides examples of MWEs in English to illustrate the wide variety among these expressions, as seen in Table 1. While they are a heterogenous family of expressions, what unifies them is a lack of compositional linguistic analysis and psycholinguistic evidence has been given for their predictability and conventionalization. Our unique approach is to adapt different computational metrics to describe the heterogeneity within these MWEs and whether it is observable at the brain level.

MWE comprehension was shown to be distinct from other kinds of language processing. For instance, it is well-established at the behavioral 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 from a psycholinguistic perspective (Clark and Wilkes-Gibbs, 1986; Clark and Marshall, 2002; Hale, 2006; Levy, 2008).This would follow if MWEs were remembered as chunks, in the sense of (Miller, 1956) that was later formalized by (Laird et al., 1986; Rosenbloom and Newell, 1987). In this study we investigate to what extent MWEs are processed as chunks or built-up compositionally during online sentence processing. By repurposing metrics which are traditionally used to identify collocations in corpus linguistics, we utilize them to investigate the different levels of compositionality within MWEs at the brain level.

Linguistic phenomena	Examples
fixed phrases	*per se*, *by and large*
noun compounds	*black coffee*, *cable car*
verb compounds	*give a presentation*, *come along*
binomials	*heaven and hell*, *safe and sound*
complex prepositions	*in spite of*
idioms	*break the ice*, *spill the beans*

Table 1: A wide variety of linguistic phenomena that are considered to be MWEs.

Earlier neuroimaging work on compositional-

[*] Co-first authors contributed equally to this work.

ity and lexical prediction by Willems et al. (2016) have addressed this issue in a broader sense using computational measures of entropy and surprisal. In natural language processing, MWEs have also been shown to have graded levels of compositionality (Salehi et al., 2015).

From a human language processing perspective, as Titone and Connine (1999) and Bhattasali et al. (2018) have discussed previously, these MWEs cannot simply be sorted into bipartite categories depending on whether they are processed as chunks or compositionally. Using the specific case of idioms, the authors in the first paper argue against an exclusively noncompositional or compositional approach and propose a hybrid approach to these expressions that ascribes noncompositional and compositional characteristics to these expressions. In a similar vein, the authors in the second paper provide neuroimaging evidence to show that these expressions fall along a graded spectrum and could be differentiated based on various aspects. Moreover, MWEs could be further distinguished based on predictability, modifiability, conventionalization, semantic opacity, among other aspects.

In this study, we utilize two Association Measures, Pointwise Mutual Information and Dice's Coefficient to capture respectively the degree of conventionalization and degree of predictability within these expressions. Furthermore, we probe whether these computational measures and their hypothesized cognitive instantiations are discernible at the cerebral level during naturalistic sentence processing.

2 Background

2.1 MWEs: A Gradient Approach

While Association Measures are commonly used in computational linguistics to identify MWEs since ngrams with higher scores are likely to be MWEs (Evert, 2008), in this study they are adapted as a gradient predictor to describe the MWEs within the text.

Krenn (2000) suggests that PMI and Dice 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 (i.e., the words in an MWE tend to co-occur frequently with each other), these two association measures were chosen to describe the strength of association between the identified word clusters (cf. identification method in Al Saied et al. (2017)).

2.1.1 Pointwise Mutual Information

The first measure we use is Pointwise Mutual Information (PMI) (Church and Hanks, 1990). 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). MWEs that receive a higher PMI score are seen as more conventionalized (Ramisch et al., 2010). Formally, PMI is a log-ratio of observed and expected counts:

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

2.1.2 Dice's Coefficient

The second measure used in this study is Dice's Coefficient (Dice, 1945; Sørensen, 1948). Dice's coefficient is used to identify rigid MWEs with strong association (Evert, 2008; Smadja et al., 1996). It is the ratio of the frequency of the sequence over the sum of the unigram frequency of the words in the sequence. E.g., for a bigram the two ratios are averaged by calculating their harmonic mean. The harmonic mean only assumes a value close to 1 (the largest possible Dice score) if there is a strong *prediction* in both directions, from `w1` to `w2` and vice versa. The association score will be much lower if the relation between the two words is asymmetrical.

This measure takes into account the length of the MWEs and the value ranges between 0 and 1:

$$\text{Dice} = \frac{n \times c(w_n^1)}{\Sigma_{i=1}^n c(w_i)} \qquad (2)$$

A higher value for the Dice Coefficient indicates that the two tokens do not occur together by chance. While PMI is systematically higher at the end of a word cluster Dice is not. Since Dice coefficient focuses on cases of very strong association rather than the comparison with independence as PMI does, it can be interpreted as a measure of predictability (Evert, 2008). Moreover, compared to PMI, Dice coefficient captures words co-occurrence in a certain order.

2.2 Association Measures as a Cognitively Plausible Metric

While earlier work has focused on individual types of MWEs, this study investigates the cognitive processes underlying the comprehension of heterogeneous MWEs differing along the lexical association of the words that compose them. Specifically, it is hypothesized that different association measures would map onto different cognitive aspects of MWEs, such as how predictable they are, how cohesive they are, how conventionalized they are, how frozen they are etc.

MWE	PMI	Dice
boa constrictor	7.935	10
fairy tale	6.165	6.422
coloured pencil	6.545	1.926
heart skipped a beat	10	0.001
gesture of weariness	5.125	0.001
object of curiosity	5.096	0.001
a dirty trick	5.603	0.001
united states	1.859	0.005
against all odds	6.012	0.013
sense of urgency	6.255	0.004
christmas tree	4.485	1.233
good morning	3.783	1.433
find out	3.479	1.240
come into	3.067	0.683

Table 2: Example of MWEs with two Association Measures: Pointwise Mutual Information and Dice's Coeffecient. Values highlighted in dark green indicate high scores while values highlighted in light green indicate low scores.

Thus, these association measures are used and adapted to describe different facets of MWEs. As presented above, PMI is taken to quantify the degree of conventionalization within these MWEs (Ramisch et al., 2010). Dice is taken to represent the degree of predictability of these MWEs (Evert, 2008). In Table 2, we can compare these measures on a set of identified word clusters. For example, expressions like *object of curiosity*, *gesture of weariness*, and *heart skipped a beat* would be considered highly conventionalized given their high PMI score but less predictable, given their low Dice score. As per these metrics, both *boa constrictor* and *fairy tales* are highly conventionalized and highly predictable whereas expressions like *united states* and *come into* are neither highly conventionalized nor highly predictable.

If we visually compare these scores for all 669 unique MWEs, as in Figure 1 below, we can also notice an interesting pattern. The values for PMI are spread across the axis and thus, the expressions are along a graded spectrum of conventionalized and have more fine-grained distinctions. On the other hand, since Dice is used to identify rigid MWEs, it tends to cluster the expressions around each end of the spectrum. We interpret these two different distributions of variance as enabling us to model different cerebral activation patterns of lexical association in MWEs processing at the brain level. Thus we repurpose Dice and PMI to represent different ongoing lexical processes.

Wiechmann (2008) also gave a cognitive dimension to the idea of association measures in order to investigate the association between a verb and its syntactic frames. He evaluated the measures against how well it could predict human reading behavior in an eye-tracking study. Our approach is similar to Wiechmann's cognitive-oriented approach since we also compare different association measures and test it against neural data, instead of behavioral data. An earlier study by Bhattasali et al. (2018) has illustrated how PMI specifically can be used to show not only the graded spectrum of compositionality within MWEs, but also how the more cohesive expressions implicate memory-related areas whereas the less cohesive expressions implicate well-known syntactic structure-building areas.

3 fMRI Study

3.1 Method

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 (cf. Table 3).

3.2 Stimuli & MWE Identification

The English 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 19,171 tokens; 15,388

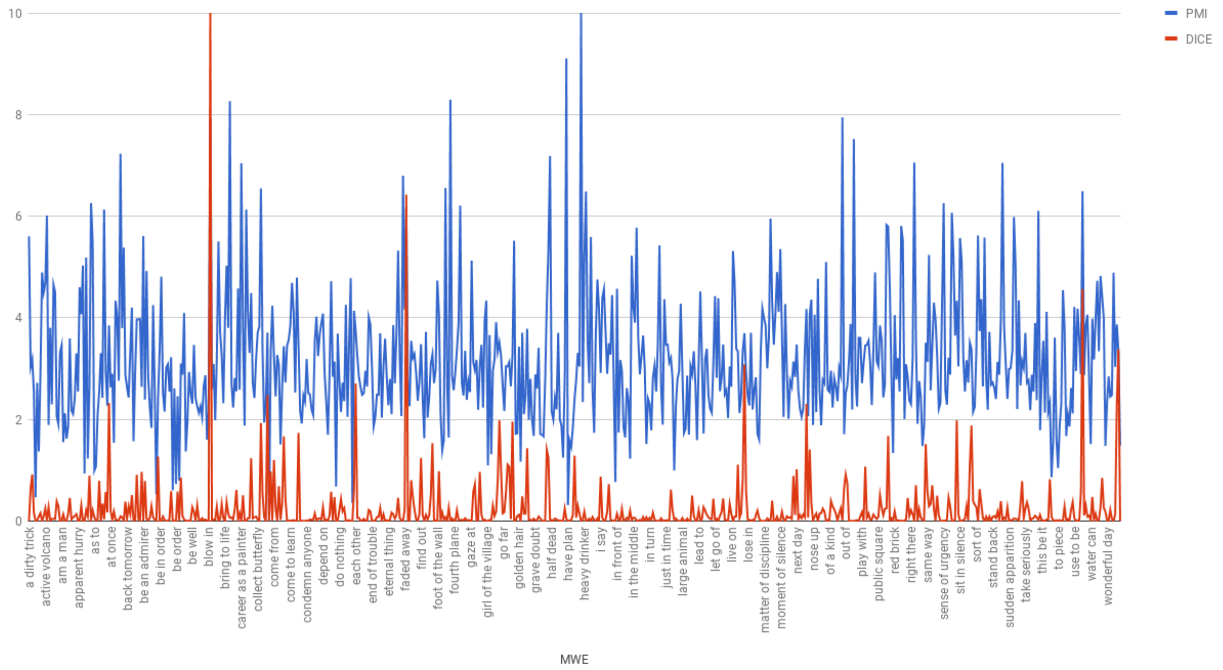

Figure 1: Comparing Pointwise Mutual Information (in blue) with Dice's Coefficient (in red); the former illustrates more fine-grained gradience; scaled up for visual purposes

words and 1,388 sentences, and lasting over an hour and a half.

Within this text, 669 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. The analyzer was trained on examples from the Children's Book Test (CBT) from the Facebook bAbI project (Hill et al., 2015) 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 also 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).

3.3 Participants

56 participants were scanned and 5 of them were excluded since they had incomplete scanning sessions. Participants included 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.

3.4 Presentation

Participants listened to the entire audiobook for 1 hour and 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.

3.5 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.

4 Data Analysis

4.1 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).

4.2 Statistical Analysis

The research questions presented above in section 2 motivates a statistical analysis that performs a comparison where fMRI signal is modeled in two General Linear Models (GLM) : one by Dice scores tagged on the identified MWEs (Model 2) versus one where PMI scores are quantifying the conventionality of each MWE in the Little Prince (Model 1).

fMRI data were analyzed in the following way: for each subject, and at each brain location (voxel), the time course of activation was submitted to a multiple linear regression that estimated the specific effect of each predictor (cf. 4.2.1), after convolution by a standard hemodynamic response (Poldrack et al., 2011).

The effects of the predictors - the increase in r^2 associated to them - were then submitted to second level analyses to test for significance at the group level. Model comparisons using root-means square (r^2) maps was carried out using a Python pipeline in order to evaluate the goodness of fit of the two Association Measures with BOLD signal (cf. 4.2.2).

4.2.1 GLM Analyses: Single-subject statistics

At the single-subject level, the observed timecourse of the brain's hemodynamic response (BOLD - Blood Oxygenation Level Dependent) in each voxel was modeled by the predictors in Table 3 including one of the two Association Measures under analysis calculated as illustrated in formulas given in 2.1), and time-locked at the offset of each word or MWE in the audio-book*.

The predictors shown in Table 3 were convolved using SPM's canonical HRF (Hemodynamic Response Function, Friston et al. (2007)). The two neuroimaging models (i.e. with PMI or with Dice) also included four control variables (confounds) as shown in Table 3.

Model 1: with PMI 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 §3.2 and all 669 unique MWEs were annotated 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) and the formula given above in 2.1. COCA is a large, genre-balanced corpus of American English and contains contains more than 560 million words of text, equally divided among spoken, fiction, popular magazines, newspapers, and academic texts.

Additionally, we entered four regressors of non-interest into the regression analysis: word offset, word frequency (Brysbaert and New, 2009), pitch, intensity which serve to improve the sensitivity, specificity and validity of activation maps (Bullmore et al., 1999; Lund et al., 2006). These predictors were added to ensure that conclusions about MWE processing would be specific to the cognitive processes they were taken to instantiate, as opposed to more general aspects of speech perception. Specifically, lexical frequency of each word was added as a covariate of non-interest, to statistically factor out effects of general word frequency, that may correlate with other types of

*For more details about the hemodynamic response, please see chapter 2 of Kemmerer (2014).

Predictors	Description
Association Measure PMI or DICE	Word-by-word on MWEs (§2.1)
Word rate	Tags the offset of each spoken word in time
Word frequency	Word-by-word log-frequency in movie subtitles
F0	Fundamental frequency of the narrator's voice, which reflects pitch
RMS amplitude	Root Mean Square Amplitude of the narrators voice, which reflects intensity, an acoustic correlate of volume

Table 3: Predictors used in the fMRI Analysis.

expectations. 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.

Model 2: with Dice Model 2 is similar to Model 1 and uses the same predictors. However, instead of PMI scores, the MWEs were annotated with their corresponding Dice's coefficient scores. These were also calculated using corpus frequency counts from COCA and the `mwetoolkit`.

4.2.2 r^2 Model comparison

The research questions presented above in section 2 motivates a statistical analysis that performs a comparison where fMRI signal on MWEs is modeled in the above presented GLMs by PMI versus Dice measures.

r^2 model comparison For every subject, we compute how much the inclusion of each variable of interest (i.e. Dice and PMI) increases the cross-validated r^2. Hence, the r^2 scores represent the variance explained in each voxel by the variable instantiating the MWE processing Dice or PMI respectively provide.

Group-level statistics To compare the impact of the two variables on fMRI signal explanation (i.e. r^2 increase of each variable), we performed a paired t-test on each individual r^2 brain map, and obtained the map in Figure 2 showing where one of the variables explains significantly better the signal than the other (see clusters on Table 4).

5 Results - Fit with fMRI signal

We performed an r^2 comparison to test which Association Measure on MWEs provided the better fit to the fMRI signal recorded during *The Little Prince*.

Dice vs. PMI The two different Association Measure were tested (Dice and PMI), and Dice, taken to represent the degree of predictability, was shown to be the best fitting the BOLD signal of these two models. Figure 2 (clusters coordinates and statistics, cf. Table 4), shows the significance (z-scores after Bonferroni correction with p < 0.05) of the difference in r^2 scores with a cluster threshold of 10 voxels.

Of the two Association Measures , the Dice measure (i.e. degree of predictability) had a significant predictive value in well-known language areas such as temporal regions, although mainly right-lateralized.

6 Discussion

The present neuroimaging study offers a first experimental grounding to the fact that a computational measure instantiating lexical prediction has a better fit with brain activity elicited by processing MWEs in certain regions of the language network. In both anterior and posterior portions of language network - and specifically in temporal areas - this lexical knowledge based process has a significant predictive value.

This result is in line with earlier work on lexical prediction with computational measures like entropy and surprisal by Willems et al. (2016) where temporal regions were identified together with right lateralized frontal ones.

Assuming Dice operationalizes some predictive processes within complex lexical items, these predictive processes are plausibly linked to higher demands in semantic combinatorial operations, as

Regions for Dice >PMI	Cluster size	MNI Coordinates			z-scores
	(in voxels)	x	y	z	
R Superior Temporal Gyrus (BA 38)	47	48	10	-26	5.80
R Middle Temporal Gyrus	84	54	-18	-10	6.09
R Middle Temporal Gyrus (BA 22)	98	48	-36	2	5.85
R Superior Temporal Gyrus (BA 22)	70	48	-12	2	5.83
R Middle Temporal Gyrus (BA 22)	16	58	-46	2	5.14
L Superior Temporal Gyrus	13	-62	-18	6	5.64
R Superior Frontal Gyrus	10	20	56	12	5.53
R Inferior Frontal Gyrus (BA 45)	10	48	20	14	5.64
L Supramarginal Gyrus	22	-56	-56	22	5.37
R Inferior Parietal Lobule/ Superior Temporal Gyrus (BA 40)	10	62	-46	22	5.44
R Inferior Parietal Lobule/ Superior Temporal Gyrus (BA 40)	16	54	-46	22	5.45
R Superior Frontal Gyrus	35	20	42	34	5.69
R Cingulate Gyrus	17	2	-34	34	5.85
R Precenus	22	32	-72	36	5.76
L Inferior Parietal Lobule	12	-34	-58	46	5.17

Table 4: Significant clusters for Dice's Coefficient versus Pointwise Mutual Information after Bonferroni correction with $p < 0.05$, based on R2 analysis in §4.2.2, and shown in Figure 2

.

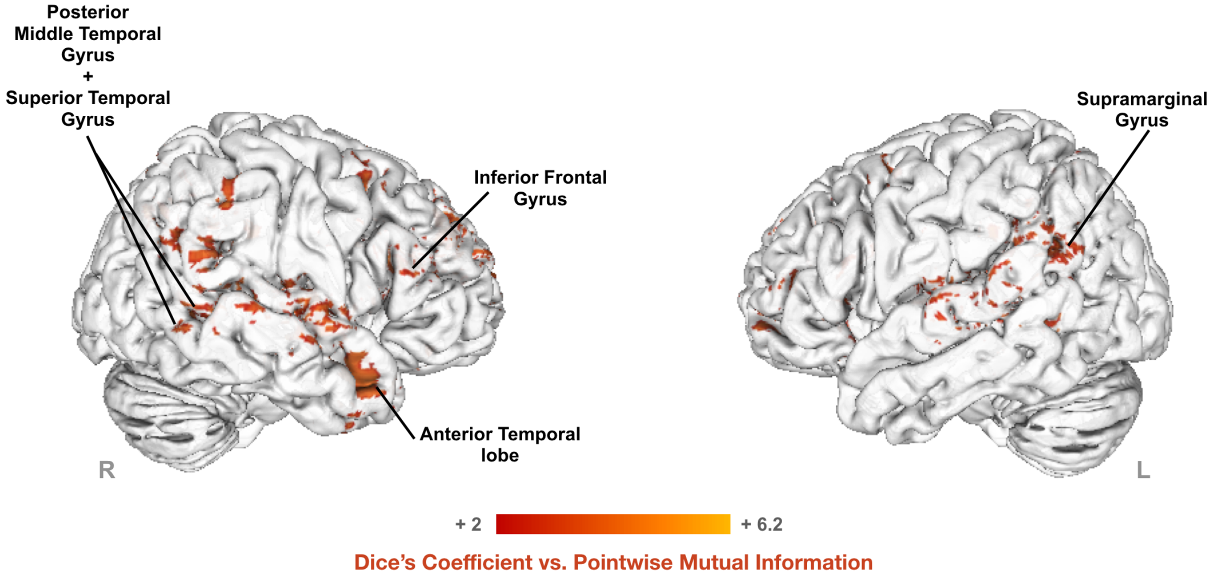

Figure 2: Z-map showing regions having a significant effect for Dice's coefficient versus Pointwise Mutual Information after Bonferroni correction with $p < 0.05$

reported in previous neuroimaging studies investigating semantic combinatorial processes through comparing meaningful and less meaningful word combinations (Price et al., 2015; Graves et al., 2010). Crucially, the graded psycholinguistic measures about lexical combination tested in these studies elicit similar areas as the regions where a better fit to the fMRI signal is observed in the present study.

Based on the formula, Dice helped us to factor out effects of length in longer MWEs and provided us with a more abstract measure given its bidirectional association. This could be a reason that it was a better fit to the BOLD signal, compared to PMI which is biased based on the length of the expression.

Lastly, Dice's Coefficient is a more rigid measure of lexical association compared to Pointwise Mutual Information, as seen in Fig. 1. Hence, Dice clusters highly predictable expressions versus less predictable ones, giving rise to two main groups. PMI displays more fine-grained distinctions overall (compared to Dice) and thus, captures the spectrum of compositional gradience within these MWEs as shown in a previous neuroimaging study. Bhattasali et al. (2018) showed that increasing values of PMI activates the network of syntactic building. However, the fact that Dice is the better fit between the two is interesting since it suggests that a bimodal distribution of gradience is cognitively more plausible than a fine-tuned approach to gradience, specifically in posterior temporal areas. Thus, this paves the way for further investigations regarding which computational measures are more cognitively pertinent to grasp a better understanding of human cognition and its neural substrates.

7 Conclusion & Further Work

Overall, this study examines MWEs through the lens of two different Association Measures, Pointwise Mutual Information and Dice's Coefficent. We investigate to what extent these computational measures, operationalizing conventionalization and predictability, and their underlying cognitive processes are observable during on-line sentence processing. Our results show that Dice's Coefficient, formalizing the degree of predictability, is a better predictor of cerebral activation for processing MWEs and this suggests it is a more cognitively plausible computational metric in temporal areas where previous neuroimaging literature identified lexical predictive processes.

Apart from Association Measures, a future approach would be to investigate different metrics to capture other nuances between these MWEs. 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).

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1607441 and by the French National Research Agency (ANR) under grant ANR-14-CERA-0001. This research has also been supported in part by ONR MURI Award N00014–18–1–2670.

We are grateful to Wenming Luh for help in preprocessing, to Hazem Al Saied and Matthieu Constant for help in identifying MWEs, and to Eric de la Clergerie for valuable help in interpreting Association Measures. We also thank the anonymous reviewers for their detailed comments on the original manuscript.

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. Association for Computational Linguistics.

Shohini Bhattasali, Murielle Fabre, and John Hale. 2018. Processing MWEs: Neurocognitive bases of verbal MWEs and lexical cohesiveness within MWEs. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 6–17.

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

Marc Brysbaert and Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods*, 41(4):977–990.

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.

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

Herbert H Clark and Catherine R Marshall. 2002. Definite reference and mutual knowledge. *Psycholinguistics: critical concepts in psychology*, 414.

Herbert H Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1–39.

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

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. *Corpus linguistics. An international handbook*, 2:1212–1248.

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.

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.

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*.

David Kemmerer. 2014. *Cognitive neuroscience of language*. Psychology Press.

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.

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

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

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.

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.

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

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. 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.

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

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.

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.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 977–983.

Anna Siyanova-Chanturia. 2013. Eye-tracking and ERPs in multi-word expression research: A state-of-the-art review of the method and findings. *The Mental Lexicon*, 8(2):245–268.

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

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

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.

Debra A Titone and Cynthia M Connine. 1999. On the compositional and noncompositional nature of idiomatic expressions. *Journal of pragmatics*, 31(12):1655–1674.

Daniel Wiechmann. 2008. On the computation of collostruction strength: Testing measures of association as expressions of lexical bias. *Corpus Linguistics and Linguistic Theory*, 4(2):253–290.

Roel M Willems, Stefan L Frank, Annabel D Nijhof, Peter Hagoort, and Antal Van den Bosch. 2016. Prediction during natural language comprehension. *Cerebral Cortex*, 26(6):2506–2516.