Measuring Cross-Lingual Semantic Similarity Across European Languages

Lütfi Kerem Şenel[1,2,3], Veysel Yücesoy[1], Aykut Koç[1], Tolga Çukur[2,3,4]
[1]ASELSAN Research Center, Ankara, Turkey
[2]Department of Electrical and Electronics Engineering, Bilkent University, Ankara, Turkey
[3]Sabuncu Brain Research Center, UMRAM, Bilkent University, Ankara, Turkey
[4]Neuroscience Program, Bilkent University, Ankara, Turkey
Email: {lksenel,vyucesoy,aykutkoc}@aselsan.com.tr, cukur@ee.bilkent.edu.tr

***Abstract*—This paper studies cross-lingual semantic similarity (CLSS) between five European languages (i.e. English, French, German, Spanish and Italian) via unsupervised word embeddings from a cross-lingual lexicon. The vocabulary in each language is projected onto a separate high-dimensional vector space, and these vector spaces are then compared using several different distance measures (i.e., correlation, cosine etc.) to measure their pairwise semantic similarities between these languages. A substantial degree of similarity is observed between the vector spaces learned from corpora of the European languages. Null hypothesis testing and bootstrap methods (by resampling without replacement) are utilized to verify the results.**

***Keywords*—language models; cross-lingual semantic similarity; natural language processing; semantic similarity; word embedding**

I. Introduction

Language similarity has been studied by researchers from different domains using numerous statistical, linguistic and neuroscientific approaches. [1] and [2] tried to fit linear and probabilistic models for semantic similarity between languages in order to achieve word to word translation. In another study, [3] tried to measure semantic similarities between the words within a morphologically rich language (such as German or Greek) using co-occurrence and context based similarity metrics. In [4], lexical similarities (similarity in both form and meaning) across different languages have been studied from a linguistic perspective. In an application to the neuroscientific domain, [5] examined the similarity of semantic representations of same-meaning words in first and second languages of bilingual subjects. An important implication of this previous study was that semantic structure is mostly shared across different languages although individual words might possess different acoustic properties. Semantic properties of languages are commonly assessed via word embedding that is a mathematical representation of words that projects linguistic vocabulary onto a vector space of predetermined number of dimensions where semantic relations of the words are preserved. Idea of representing words by real vectors dates back to early studies that have proposed the LSA and randomized embedding techniques [6]. Since then, there have been extensive efforts to optimize word vectors for particular levels of linguistic features such as syntax, morphology or semantics [7]–[9]. The utility of such computationally derived word vectors have been amply demonstrated in neural language models in literature [10]–[12].

Two foundational works, [13] and [14], presented the impressive results by achieving unsupervised learning of vector spaces from large corpora. The learned embeddings then enable successful results on analogy tests that require capturing of relationships based on word meaning (i.e. $man - king + woman \;\; =? \; queen$).

Due to the initial success of abovementioned approaches [13], [14], these embeddings are recently forming a baseline for natural language processing (NLP) related tasks as well as some other disciplines which make use of NLP, including neuroscience, computer vision and computational linguistics. A common computer vision task that makes use of word embeddings is captioning. The aim of captioning is to form a suitable sentence that explains a given image [15]. Another example from neuroscience, which is built on the semantic properties of the words, is the study of [16] that analyzed changes in brain activity evoked by words in spoken narratives and mapped brain regions that represent various types of semantic information.

Cross-lingual semantic similarity has been studied by several researches for different natural language processing applications such as word-to-word translation [2], comparing articles from different languages [17] or finding the semantic similarity of words from different languages [18], [19]. However, to the best of our knowledge, cross-lingual semantic similarity of different languages has not been studied using more recent, powerful word embedding methods. This paper aims to study the cross-lingual semantic similarity between language varieties over a cross-lingual lexicon using word embeddings generated by the Glove [14] algorithm. In order to quantify the cross-lingual semantic similairities, representational similarity analysis (RSA) is utilized. The languages studied within the scope of this paper are five common European languages, English, French, German, Italian and Spanish.

This work was supported in part by a Marie Curie Actions Career Integration Grant (PCIG13-GA-2013-618101), by a European Molecular Biology Organization Installation Grant (IG 3028), and by a TUBA GEBIP fellowship awarded to T. ukur.

978-1-5090-3982-1/17/$31.00 ©2017 IEEE

The paper is organized as follows: Section II explains the details of the datasets that are used to generate the word embeddings for each language we study, together with all of the preprocessing steps. Selection of the vocabulary and main experimental setup are discussed in Section III. The calculation method of semantic similarity, numerical results and statistical analysis take place also in Section III. Section IV concludes the paper.

II. Datasets and Preprocessing

A. Datasets

Entire Wikipedia[1] pages in each language are selected as source for dataset since they provide a large enough and grammatically reasonable source covering a wide range of different subjects. Having compatible datasets for studying similarities between different languages is another advantage of using Wikipedia.

TABLE I. Details of corpora used for each language. Number of words and file sizes are calculated after preprocessing.

Language	# of words (M)	Size in Gb	File Date
English	1170	8,6	02.11.2016
French	295	2,4	02.11.2016
German	410	3,7	21.10.2016
Italian	220	1,8	20.11.2016
Spanish	251	2,1	20.10.2016

Latest available versions of the contents of the Wikipedia for the languages we studied are downloaded. Details of the used corpora are given in Table I. In this table, first two columns represent the features of the processed corpora and details of the preprocessing are explained in Section II-B.

B. Preprocessing

It is a common practice to use a dedicated Python script[2] to extract, clean and store text from downloaded Wikipedia database dumps. This process results in a number of files of similar size where each file contains several documents in the format shown below:

$$\boxed{< docid = “\ldots” \; url = “\ldots” \; title = “\ldots” > \cdots < /doc >}$$

After the extraction, resulting files are combined, headers and all non-alphabetic characters (punctuations, digits, etc.) are removed. Stop words which are listed by natural language toolkit (NLTK)[3] are also removed due to the assumption that they do not carry significant information about the semantic structure of a language. Then all letters are converted to lower case and finally all the remaining words are written to a text file in a single line in order to prepare a valid input for word embeddings [14]. Letters coming after apostrophes are taken as separate words (she'll becomes she ll). File sizes and the number of words contained in each dataset after the preprocessing as well as the last update dates of the Wikipedia dumps before their use in this study are given in Table I for five languages.

TABLE II. Parameters for GloVe.

Parameter Name	Value	Applied Language
VECTOR_SIZE	300	All
MAX_ITER	50	All
WINDOW_SIZE	15	All
X_MAX	100	All
VOCAB_MIN_COUNT	50	English
	17	French
	52	German
	12	Italian
	15	Spanish

Remark: The experiments described below were conducted without eliminating stop words from the corpus (datasets) as well. Resulting correlations are slightly less than the ones presented in this paper for all languages which supports the assumption that these words do not provide significant information.

III. Experiments and Results

A. Construction of Vocabulary

Word embeddings for five languages are obtained by training GloVe [14] with preprocessed Wikipedia dumps using the parameters displayed in Table II. VOCAB_MIN_COUNT which is required minimum number of occurrences for a word in a corpus in order to be used in word embedding is adjusted across languages so that all five languages have similar vocabulary sizes.

A fixed vocabulary needed to be constructed for measuring semantic similarities between languages. For a vocabulary to be able to represent a language, its size should be large enough, included words should be commonly used and it should cover a large variety of topics in order to prevent bias and to capture as many semantic relations as possible. By taking these requirements into consideration, a vocabulary is constructed by combining words from four online sources for English. Online sources that are used to construct this vocabulary and their content can be listed as follows:

1. 1000 basic words[4]
2. 1000 most frequent words[5]
3. 1000 most frequent verbs[6]
4. Categorized word list[7]

The last source, where English words are separated into many categories, is used to broaden the extent of the words.

[1] https://www.wikipedia.org

[2] http://github.com/attardi/wikiextractor

[3] www.nltk.org

[4] https://simple.wikipedia.org/wiki/ Wikipedia:List_of_1000_basic_words

[5] http://www.ef.com/english-resources/english-vocabulary/ top-1000-words

[6] http://www.talkenglish.com/vocabulary/top-1000-verbs.aspx

[7] http://www.manythings.org/vocabulary/lists/c

Manual elimination is conducted on those categorized words in order to avoid rare words. After this manual elimination, all words from four sources are combined without duplicates to form a basis vocabulary for English. As a final step, stop words which take place within NLTK are removed from all vocabularies. Word count for the combined vocabulary was 2443 just after the removal of stop words. In order to form a basis vocabulary for other languages, resulting English vocabulary is translated. A professional translation service is utilized to have reliable vocabularies. Words, whose translations cannot be expressed as a single word in at least one of the languages are eliminated from all vocabularies decreasing the size of the final vocabularies to 2107. The final vocabulary lists will be publicly available as a vocabulary dataset[8].

B. Experiment

Representational similarity analysis (RSA) is one of the techniques used for relating computational models to measured neural activities, see references [20], [21] for details, which characterizes a computational model by a distance matrix named representational dissimilarity matrix (RDM). In this study RDM is used to measure semantic similarities between different languages which are modelled through word embeddings. Following assumptions are made for finding cross-lingual semantic similarity between two languages in this study.

Assumption 1. *Assume that there are two languages L_1 and L_2. Take M words from L_1 and define $W_1 = \{w_{1,1}, w_{1,2}, \dots, w_{1,M}\}$, take M words from L_2 and define $W_2 = \{w_{2,1}, w_{2,2}, \dots, w_{2,M}\}$ where $w_{2,i}$ is word to word translation of $w_{1,i}$ for all $i \in \{1, 2, \dots, M\}$.*

Assumption 2. *Assume that there are two sets of word vectors E and S such that each set has N dimensional word vectors corresponding to words in W_1 and W_2, respectively (i.e. $E = \{e_1, e_2, \dots, e_M\}$, $S = \{s_1, s_2, \dots, s_M\}$ where $e_i \in \mathbb{R}^{N\times 1}$ is a word embedding for $w_{1,i}$ and $s_i \in \mathbb{R}^{N\times 1}$ is a word embedding for $w_{2,i}$, $i \in \{1, 2, \dots, M\}$).*

Two different metrics are used to obtain dissimilarity matrix for a language. Correlation based dissimilarity matrix $\mathsf{LE^c}$ is defined as

$$\mathsf{LE^c}_{i,j} = 1 - \frac{\hat{e}_i^T \hat{e}_j}{\|\hat{e}_i\|_2 \|\hat{e}_j\|_2} \tag{1}$$

where $\hat{e}_i$ is given by

$$\hat{e}_i = e_i - \mu_{e_i} \quad \text{and} \quad \mu_{e_i} = \frac{1}{N}\sum_{l=1}^{N} e_i[l] \tag{2}$$

and $e_i[l]$ is the l^{th} element of vector e_i.

Cosine distance based dissimilarity matrices $\mathsf{LE^d}$ is defined as

$$\mathsf{LE^d}_{i,j} = 1 - \frac{e_i^T e_j}{\|e_i\|_2 \|e_j\|_2} \tag{3}$$

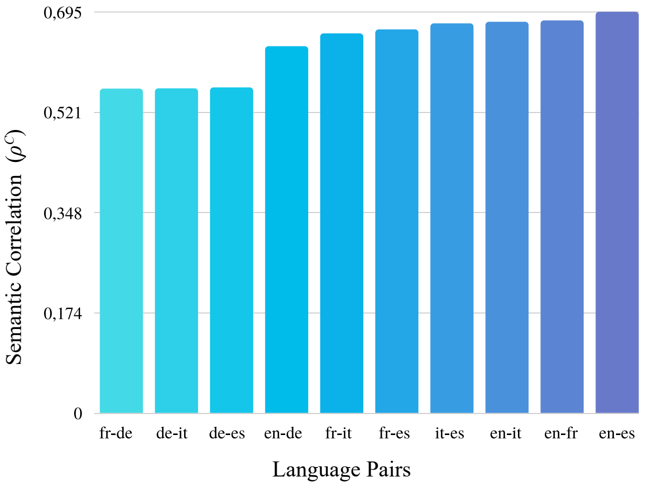

Fig. 1. Semantic similarities of European language pairs. Height of the columns represent the calculated semantic correlation between the corresponding language pair. ISO 639-1 two letter language codes are used for language names (en:English, de:German, fr:French, it:Italian, es:Spanish).

Dissimilarity matrices $\mathsf{LS^c}$ and $\mathsf{LS^d}$ for set S are defined in the same manner.

Then semantic similarity between languages L_1 and L_2 can be calculated for two different metrics as

$$\rho^c = corr(\mathsf{LE^c}, \mathsf{LS^c}), \quad \rho^d = corr(\mathsf{LE^d}, \mathsf{LS^d}) \tag{4}$$

where correlation is calculated over only off-diagonal entries of the dissimilarity matrices.

Equation 4 represents the core comparison method of this paper. As usual, ρ^c or ρ^d value of near 0 means a weak relation whereas a value near 1 means a strong semantic similarity between a language pair. Each language pair is prepared according to the explanations in Section II and Assumptions 1,2. For each pair, the experiment is conducted twice (once for each parameter in Equation 4) and the results for ρ^c are depicted in Table III in the column 'Semantic Similarity' and also in Fig. 1 for a better visual comparison.

C. Statistical Testing

Since the experiment results give only an estimate of the unknown similarity parameter, the statistical reliability of this estimate is of question. In order to test the reliability of the estimate and to impose some confidence interval, different statistical procedures are applied.

1) Bootstrap (resampling without replacement): Bootstrap is a general term that defines a set of resampling methods to find statistical variance estimation. There are two basic approaches to utilize bootstrap: resampling with and without replacement. In this paper, resampling without replacement is used. More precisely, it is possible to express the method as follows: Assume that Assumptions 1 and 2 hold. Select $b < M$ and choose b vectors from E and S randomly at each iteration k (i.e. define b random indices without repetition $T = \{t_1, t_2, \dots, t_b\}$ where $t_i < M$ for all $i \in \{1, 2, \dots, b\}$.

[8] Send an email to the first author to get the dataset

TABLE III. SEMANTIC SIMILARITIES OF LANGUAGE PAIRS CALCULATED USING CORRELATION DISTANCE METRIC, CORRESPONDING CONFIDENCE INTERVALS CALCULATED FROM BOOTSTRAP WITHOUT REPLACEMENT, MEAN AND STANDARD DEVIATIONS SEMANTIC SIMILARITIES OVER 1000 BOOTSTRAP ITERATIONS AND p VALUES FOR TWO NULL HYPOTHESIS TESTS.

Language	**Semantic Similarity**	**Confidence Interval (95%)**	**Bootstrap (μ and σ)**	**NHT1**	**NHT2**
English - French	0.6801	0.6688 - 0.6914	0.6799 0.0057	$p <$ 0.001	$p <$ 0.001
English - German	0.6354	0.6264 - 0.6445	0.6353 0.0046		
English - Italian	0.6777	0.6681 - 0.6879	0.6779 0.0050		
English - Spanish	0.6951	0.6861 - 0.7048	0.6950 0.0048		
French - German	0.5619	0.5502 - 0.5743	0.5619 0.0063		
French - Italian	0.6576	0.6462 - 0.6702	0.6577 0.0062		
French - Spanish	0.6645	0.6527 - 0.6772	0.6642 0.0062		
German - Italian	0.5624	0.5519 - 0.5743	0.5628 0.0056		
German - Spanish	0.5640	0.5528 - 0.5750	0.5638 0.0055		
Italian - Spanish	0.6750	0.6642 - 0.6850	0.6744 0.0056		

Then form $X = \{e_{t_1}, e_{t_2}, \dots, e_{t_b}\} = \{x_1, x_2, \dots, x_b\}$ and $Y = \{s_{t_1}, s_{t_2}, \dots, s_{t_b}\} = \{y_1, y_2, \dots, y_b\}$). Define a new correlation based distance matrix as

$$\mathsf{LE}^{(k)}_{i,j} = 1 - \frac{\hat{x}_i^T \hat{x}_j}{\|\hat{x}_i\|_2 \|\hat{x}_j\|_2} \tag{5}$$

where $i, j \in \{1, 2, \dots, b\}$ for $k \in \{1, 2, \dots, r\}$. $\mathsf{LS}^{(k)}$ is calculated similarly.

Note that $\mathsf{LE}^{(k)}$ and $\mathsf{LS}^{(k)}$ are matrices of size $b \times b$ and r is the experiment repetition number. Then calculate the correlation c_k between $\mathsf{LE}^{(k)}$ and $\mathsf{LS}^{(k)}$ as explained in Section III. Resulting mean and standard deviation values of c_k's for $r = 1000$ iterations and $b = 0.8M$ are presented in Table III along with the 95% confidence intervals for the correlation metric. Lower and upper bounds of the confidence intervals are chosen as the 25^{th} and 975^{th} similarity values coming from 1000 bootstrap iterations after sorting them in ascending order.

2) Null Hypothesis Testing: In order to confirm the validity of the results, null hypothesis is tested by randomizing the process at two different levels. In the first test (NHT1), instead of the lists W_1 and W_2 which are word to word translations of each other, new lists are constructed from random shuffles of W_1 and W_2 breaking apart the semantic coherence between word pairs among languages. Their word vectors are used to find RDM's, and finally correlations are obtained using the same procedure.

In the second test (NHT2), dimensions of the word vectors in E and S are shuffled independent of other words in each set so that the projections of the words onto vector space are partly randomized breaking the semantic relations between words in a vector space. RDM's are constructed from resulting word vector sets and correlations are obtained from RDM's in the same manner.

Both procedures are repeated 1000 times for each language pair and p values are calculated as the probability of having a correlation greater than the mean of bootstrap trials which are nearly identical with corresponding semantic similarities. Resulting p values are presented in Table III. Null hypotheses are rejected for all language pairs, for both tests.

D. Results

It is important to note that, although two different metrics are discussed in this paper, results are presented only for correlation distance in Table III. This is because similarities calculated using cosine distance metric are nearly identical with the ones presented in Table III.

Table III yields a significant cross-lingual semantic similarity between the five European languages. One of the most prominent results is the relatively low similarity between German and Romance (Latin originated) languages namely French, Italian, Spanish, where average similarity is 0.5628, compared to similarity between German and English which is calculated as 0.6354. This is understandable due to the different origins of separate language groups (Romance and Germanic). English is originally from the same group with German (Germanic), therefore these two languages have expectedly higher similarity compared to that among German and Romance languages. However, modern English is a result of many historical interactions with Romance languages demonstrated by the fact that most of English vocabulary is borrowed from these languages. This explains the relatively high similarity between English and Romance languages. Average correlation among Romance languages, which is 0.6657, is greater than the correlation between Germanic languages (German, English), and this result could be due to the change in semantic structure of English as a result of its interactions with Romance languages.

Bootstrapping and null hypothesis testing verify that obtained correlations are due to the similar semantic structure of the languages. Both null hypothesises explained in Section III-C2 are rejected for all language pairs since $p < 0.001$. With these procedures, it is demonstrated that obtained correlations are not due to inherent structure of the word vectors but the word to word semantic similarities between languages.

IV. Discussion

Here cross-lingual semantic similarities between five European languages (i.e. English, French, German, Italian and Spanish) are quantified using representational similarity analysis. Cross-lingual semantic similarity between language pairs was defined by means of word embeddings over a fixed cross-lingual lexicon. This dictionary was generated in English from different word lists to include as many different word categories as possible. It was translated to other languages manually by professionals to form a reliable basis for all languages.

Cross-lingual semantic similarity is a measure of similarity between languages which deeply joins the temporal relations and co-occurrences of the word pairs. As a result of this feature, in this paper it is suggested that these embeddings are well suited for multilingual tasks such as word to word translation, neuroscience studies on bilingual speakers, multilingual text classification, etc. For example a possible application of semantic similarity between word embeddings is to classify the subjects of texts from different languages. Assume that there exist a dataset of texts in different languages about a broad range of topics. A subject classifier can be trained for all languages using the semantic similarity between word embeddings to improve the predictions of the subject of a single text.

In this paper we investigated five common European languages that are pervasively used in Europe, and quantified their semantic similarity via a representational similarity analysis of word embeddings obtained for each language independently. In future work, we plan to examine a broader set of languages from distinct families to obtain more comprehensive evaluations.

References

[1] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation," *arXiv preprint arXiv:1309.4168*, 2013.

[2] I. Vulic and M.-F. Moens, "Probabilistic models of cross-lingual semantic similarity in context based on latent cross-lingual concepts induced from comparable data," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*. ACL, 2014, pp. 349–362.

[3] K. Zervanou, E. Iosif, and A. Potamianos, "Word semantic similarity for morphologically rich languages." in *LREC*, 2014, pp. 1642–1648.

[4] G. F. Simons and D. F. Charles, *Ethnologue: Languages of the World*. SIL International, 2017. [Online]. Available: http://www.ethnologue.com

[5] J. Correia, E. Formisano, G. Valente, L. Hausfeld, B. Jansma, and M. Bonte, "Brain-based translation: fmri decoding of spoken words in bilinguals reveals language-independent semantic representations in anterior temporal lobe," *The Journal of Neuroscience*, vol. 34, no. 1, pp. 332–338, 2014.

[6] D. Ravichandran, P. Pantel, and E. Hovy, "Randomized algorithms and nlp: Using locality sensitive hash function for high speed noun clustering," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2005, pp. 622–629.

[7] A. Mnih and G. Hinton, "Three new graphical models for statistical language modelling," in *Proceedings of the 24th International Conference on Machine Learning*. ACM, 2007, pp. 641–648.

[8] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, vol. 1631. Citeseer, 2013, p. 1642.

[9] T. Luong, R. Socher, and C. D. Manning, "Better word representations with recursive neural networks for morphology." in *CoNLL*, 2013, pp. 104–113.

[10] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of Machine Learning Research*, vol. 3, no. Feb, pp. 1137–1155, 2003.

[11] J. Turian, L. Ratinov, and Y. Bengio, "Word representations: A simple and general method for semi-supervised learning," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 384–394.

[12] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th International Conference on Machine Learning*. ACM, 2008, pp. 160–167.

[13] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[14] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162

[15] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137.

[16] A. G. Huth, W. A. de Heer, T. L. Griffiths, F. E. Theunissen, and J. L. Gallant, "Natural speech reveals the semantic maps that tile human cerebral cortex," *Nature*, vol. 532, no. 7600, pp. 453–458, 2016.

[17] M. Saad, D. Langlois, and K. Smaïli, "Cross-lingual semantic similarity measure for comparable articles," in *International Conference on Natural Language Processing*. Springer, 2014, pp. 105–115.

[18] I. Vulic and M.-F. Moens, "Cross-lingual semantic similarity of words as the similarity of their semantic word responses," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*. ACL, 2013, pp. 106–116.

[19] L. Dai and H. Huang, "An english-chinese cross-lingual word semantic similarity measure exploring attributes and relations." in *PACLIC*, 2011, pp. 467–476.

[20] H. Nili, C. Wingfield, A. Walther, L. Su, W. Marslen-Wilson, and N. Kriegeskorte, "A toolbox for representational similarity analysis," *PLoS Comput Biol*, vol. 10, no. 4, p. e1003553, 2014.

[21] N. Kriegeskorte and R. A. Kievit, "Representational geometry: integrating cognition, computation, and the brain," *Trends in Cognitive Sciences*, vol. 17, no. 8, pp. 401–412, 2013.