Collecting and Predicting Neurocognitive Norms for Mandarin Chinese

Le Qiu, Yu-Yin Hsu, Emmanuele Chersoni
Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University,
11 Yuk Choi Road, Hung Hom, Kowloon, Hong Kong (China)
lani.qiu@connect.polyu.hk, {yu-yin.hsu,emmanuele.chersoni}@polyu.edu.hk

Abstract

Language researchers have long assumed that concepts can be represented by sets of semantic features, and have traditionally encountered challenges in identifying a feature set that could be sufficiently general to describe the human conceptual experience in its entirety.

In the dataset of English norms presented by Binder et al. (2016), also known as **Binder norms**, they introduced a new set of *neurobiologically motivated* semantic features in which conceptual primitives were defined in terms of modalities of neural information processing. However, no comparable norms are currently available for other languages.

In our work, we built the Mandarin Chinese norm by translating the stimuli used in the original study and developed a comparable collection of human ratings for Mandarin Chinese. We also conducted some experiments on the automatic prediction of the Chinese Norms based on the word embeddings of the corresponding words to assess the feasibility of modeling experiential semantic features via corpus-based representations.

1 Introduction

A longstanding research trend in semantics assumes that the conceptual content of lexical items can be decomposed into semantic features identifying basic meaning components (Vigliocco and Vinson, 2007). Such features represent semantic primitives that can be present or absent in the semantic representation of a lexeme, such as *boy* in Example (1).

(1) *boy* [+MALE, -MATURE ...]

However, this type of view has some critical limitations: First, discrete features are not suitable to address the gradient prototypicality of feature-to-concept associations (Murphy, 2002). Second, these feature sets tend to be manually selected, and are generally tailored to a few *in vitro* examples; thus, they are unable to account for large portions of the lexicon of natural languages (Chersoni et al., 2021).

On one hand, featural representations have the advantage of *human interpretability*, as they label the dimensions of word meanings explicitly, and provide explanatory factors for their semantic behavior; for example, the similarity between *beer* and *coffee* can be explained by assuming that they share the semantic feature of LIQUID. On the other hand, this type of features is highly subjective, and can only be collected through a time-consuming process of elicitation from human subjects (e.g. McRae et al. (2005); Vinson and Vigliocco (2008); Devereux et al. (2014); Buchanan et al. (2019)).

An alternative was proposed by Binder et al. (2016) using **brain-based semantics** based on *modalities of neural information processing*. After reviewing extensive evidence from studies of human physiology, the authors proposed a dataset of 535 words described in terms of 68 experiential features, each of which was associated with a specific neural processing in the neurobiological literature. The features were categorized according to 14 different domains of experience (Table 1).

The proposal by Binder et al. (2016) should naturally extend to other languages: If the features are genuinely neurobiologically motivated, it should also be possible to use them to describe the essential meaning components of languages other than English.[1] However, to the best of our knowledge, Binder-like norms are currently only available for the English language.[2]

[1] See also the recent work of Blasi et al. (2022) on the need for cognitive science studies to look beyond English, in order to support claims of universality.

[2] A partial exception is represented by the collection of ratings published by Wang et al. (2022); see Section 2.

Domain Type	Domain	Meaning components (features)
Sensory	Vision	VISION, BRIGHT, DARK, COLOUR, PATTERN, LARGE, SMALL, MOTION, BIOMOTION, FAST, SLOW, SHAPE, COMPLEXITY, FACE, BODY
Sensory	Somatic	TOUCH, HOT, COLD, SMOOTH, ROUGH, LIGHT, HEAVY, PAIN
Sensory	Audition	AUDITION, LOUD, LOW, HIGH, SOUND, MUSIC, SPEECH
Sensory	Gustation	TASTE
Sensory	Olfaction	SMELL
Motor	Motor	HEAD,UPPER LIMB, LOWER LIMB, PRACTICE
Spatial	Spatial	LANDMARK, PATH, SCENE, NEAR, TOWARD, AWAY
Number	Number	NUMBER
Event	Temporal	TIME, DURATION, LONG, SHORT
Event	Causal	CAUSED, CONSEQUENTIAL
Event	Social	SOCIAL
Cognition	Cognition	HUMAN, COMMUNICATION, SELF, COGNITION
Evaluation	Evaluation	BENEFIT, HARM, PLEASANT, UNPLEASANT
Emotion	Emotion	HAPPY, SAD, ANGRY, DISGUSTED, FEARFUL, SURPRISED
Drive	Drive	DRIVE, NEEDS
Attention	Attention	ATTENTION, AROUSAL

Table 1: List of the domains and meaning components (features) in Binder et al. (2016).

Therefore, in our work, we adopted the same design of Binder norms: We translated the words in the Binder dataset into Mandarin Chinese, and obtained ratings from human subjects for each of the 68 Binder features per word in order to obtain a comparable dataset. Moreover, we experimented with regression algorithms to assess the extent to which such norms could be predicted automatically based on the text-derived embeddings of the corresponding words.[3]

2 Related Work

Neurosemantic decoding research, initiated by the seminal work of Mitchell et al. (2008), has the aim of creating mappings between different concept representations, typically from a corpus-derived one (such as word embedding) to one derived from human data (such as fMRI scans and semantic norms). For example, previous studies used fMRI data to learn mapping from the traditional count-based distributional models (Devereux et al., 2010; Murphy et al., 2012), including both count- and prediction-based vectors (Bulat et al., 2017; Abnar et al., 2018), and topic models (Pereira et al., 2011, 2013); the same methodology has been used to map word-embedding models onto feature (Fagarasan et al., 2015; Bulat et al., 2016; Derby et al., 2019) and modality norms (Chersoni et al., 2020) to ground the vectors in perceptual data and to make them interpretable. Due to the grounding on perceptual experience, the Binder features for English have also been used for the same purpose (Utsumi, 2018; Turton et al., 2020; Chersoni et al., 2021). Notice that, differently from property norms (McRae et al., 2005; Devereux et al., 2014), the collection process is more constrained: the properties of concepts are not freely elicited from human participants; because the Binder features are a closed set, the participants were asked to only rate the relevance of a given feature for a given concept.

We are not currently aware of any other work that has introduced Binder-like norms for languages other than English. The recent work by Wang et al. (2022) introduced a fMRI dataset for Mandarin Chinese, together with a collection of Binder ratings for the target words. However, their targets differed from those in the original study by Binder et al. (2016) (a total of 672 words from the Synonymy Thesaurus of the Harbin Institute of Technology), and the representation was limited to 54 Binder features, as some of them were excluded due to high levels of correlation with at least one of the other features. With the aim of providing a comparable and more comprehensive resource to facilitate future experiments on the prediction of crosslingual norms, we opted to retain the original set of target words and features.

3 Data Collection

Binder et al. (2016) collected ratings for 68 cognitively-motivated features for 535 words in total.[4] 242 words were selected from the Knowledge Representation in Neural Systems project (Glasgow et al., 2016), including 141 nouns, 62 verbs,

[3]Dataset and code for the experiments will be available at the following URL: `https://github.com/Laniqiu/norming`.

[4]In their paper, they used the feature label *Temperature* for features Hot and Cold, *Texture* for Smooth and Rough, and *Weight* for Light and Heavy, resulting in 65 feature categories.

Type-POS	No. of items
Concrete Objects - Nouns	275
Living Things - Nouns	126
Other Natural Objects - Nouns	19
Artifacts - Nouns	130
Concrete Events - Nouns	60
Abstract Entities - Nouns	99
Concrete Actions - Verbs	52
Abstract Actions - Verbs	5
States - Verbs	5
Abstract Properties - Adjectives	13
Physical Properties - Adjectives	26

Table 2: Concept types, parts of speech (POS), and the number of items in the dataset by Binder et al. (2016).

Word	VISION	BRIGHT	...	COGNITION	BENEFIT
公寓*(gongyu)*	5.56	3.82	...	0.86	4.60
杏子*(xingzi)*	4.24	4.06	...	1.34	3.34

Table 3: Sample of Binder vectors for the words *gongyu (apartment)* and *xingzi (apricot)*.

and 39 adjectives, while another 293 words were added to include more abstract nouns. We adopted the original set of 535 target words and 68 features proposed by Binder et al. (2016), and the original survey queries that they proposed. We translated them into Mandarin Chinese using simplified characters. This survey was used to elicit the ratings for the salience of each attribute for each target word, with the same 0-6 Likert scale used in the original study (the higher the score, the higher the relevance of a feature when one has to think about the target concept, while 0 corresponds to “feature not applicable to this concept”).

The target words and the survey queries were translated by two native speakers of Mandarin, who were Master’s students of linguistics. For features and target words, we adopted their most common and core sense in English to translate into their corresponding Chinese. While some words in colloquial uses may have multiple senses, we selected more specific words which were equally frequent to the polysemous ones and to the sense expressed by the English counterparts. We were aware that the concept of “adjective” could sometimes not easily be recognized in Chinese, just as the function of words in the *-ed* form can be ambiguous in English as either adjectival or verbal past participle. When an adjective could be interpreted as other parts of speech categories (POS), we added an adjectival suffix -的 *de* to such adjectives to avoid such potential confusion. The final version of the survey queries and the target words were manually checked by one of the authors, who is also a native Mandarin speaker. The same POS of each word were maintained for the 535 words, and each word was associated with survey questions pertaining to the 68 cognitively motivated features. One target word in the survey, *banjo*, was replaced for a more culturally relevant musical instrument, 二胡 *erhu*, while the other words were the same as their English counterparts.

As is the case for the Binder norms, we adopted a continuous rating design to obtain the attributes for each word. We collected the data on a crowdsourcing platform that is commonly used in China (问卷星 Wenjuanxing), because the rating results might occur along a continuum and could be subjective due to the speakers’ personal experiences and backgrounds, thus, a larger sample size was considered to be helpful in overcoming this issue. We obtained 8025 sets of ratings from the crowdsourcing survey; each of the 535 targets obtained 15 sets of rating results covering all 68 features. The demographics and the language backgrounds of the participants were checked before they participated in the survey. Each participant received RMB$20 after completing the survey and once their results had passed the survey’s attention checks.

After completing the survey, we measured the Spearman correlation between English and Chinese ratings. We found out that the ratings were quite consistent across languages: on average, we obtained a correlation of 0.68 across words and a correlation of 0.59 across features.

4 Experiments

In order to learn to map between word-embedding spaces and our Chinese Binder features, we trained regression models using three different regressors, namely Ridge Regression, Random Forest and Multilayer Perceptron (MLP) [5], using the ratings of the 68 features in the dataset as the dependent variables and the dimensions of pretrained word-embedding models as the independent variables.

Considering that the task requires mapping between word types that are taken out of context, we decided to use static word-embedding mod-

[5]The regression models were implemented using Scikit-learn (Pedregosa et al., 2011) with standard hyperparameters. The only exception was the MLP, for which we selected the following parameters after a parameter search: hidden_layer_sizes=(50, 10), activation='identity', early_stopping=True, max_iter=1000 (the other parameters are the default ones).

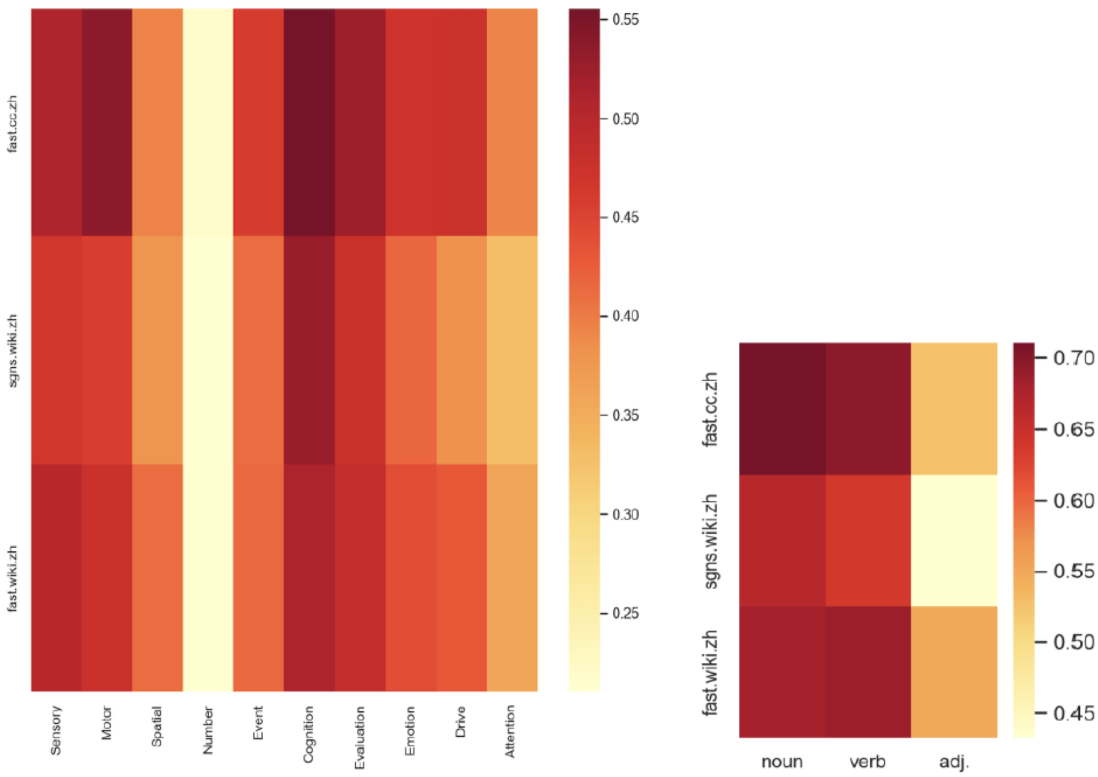

Figure 1: Feature correlation scores by domain type (left) and word correlation scores by POS (right).

els: we used four different types of embeddings: count-based sparse PPMI vectors (Church and Hanks, 1990; Bullinaria and Levy, 2007) that were trained on the Chinese Wikipedia (*ppmi.wiki.zh*; Qiu et al. (2018)), Skip-Gram vectors (Mikolov et al., 2013) that were trained on the Chinese Wikipedia (*sgns.wiki.zh*, Qiu et al. (2018)), and FastText vectors that were trained on the Chinese Common Crawl (*fast.cc.zh*) or on the Chinese Wikipedia (*fast.wiki.zh*) (Bojanowski et al., 2017)). All the embedding models had 300 dimensions as input features for the regressor, except for the sparse PPMI vectors, which had 350k dimensions. In addition, we initialized 300-dimensional random vectors for all the words in the dataset, and used them to train similar regression models as baselines (*Random*). In future, we also plan to test contextualized word embeddings (Devlin et al., 2019) in the task, although it is worth pointing out that their performance in out-of-context semantic tasks has recently been shown not to differ significantly from that of static models (Lenci et al., 2022).

Following Utsumi (2018), we adopted the **leave-one-out paradigm** for data splitting: For each of the n target words; we extracted one word out and trained a regression model on the other $n-1$ remaining words, and then we used the last word as the test set. The standard metric of the Spearman correlation was computed to compare the vectors of the Binder features predicted by the models and the gold vectors of human ratings (note that only one word was predicted for each run).

Vectors	Model	Word	Feature
fast.cc.zh	Ridge	**0.70**	**0.49**
fast.cc.zh	RandomForest	0.66	0.36
fast.cc.zh	MLP	0.69	0.40
sgns.wiki.zh	Ridge	0.66	0.44
sgns.wiki.zh	RandomForest	0.63	0.33
sgns.wiki.zh	MLP	0.66	0.38
fast.wiki.zh	Ridge	0.68	0.47
fast.wiki.zh	RandomForest	0.64	0.35
fast.wiki.zh	MLP	0.69	0.44
ppmi.wiki.zh	Ridge	0.25	0.03
ppmi.wiki.zh	RandomForest	0.50	0.07
ppmi.wiki.zh	MLP	0.15	0.03
Random	Ridge	0.26	-0.01
Random	RandomForest	0.51	-0.02
Random	MLP	0.49	0.04

Table 4: Word and Feature Spearman correlation for all regression models (top scores are in **bold**).

5 Results

The results in Table 4 reveal that embedding models based on FastText and Skip Gram had highly significant correlations with human scores, and that the FastText vectors trained on Common Crawl achieved higher scores than did any of the ones trained on Wikipedia. However, the sparse PPMI vectors had a much weaker performance, to the extent that the scores were close to the regressors initialized using the random vectors. Both the models with random and with PPMI vectors failed to achieve significant correlations at the feature level. Ridge Regression models were the most accurate, particularly for the correlations at the feature level. However, it should be said that the differences between the regressors trained with Skip-Gram and FastText are small and not significant, also due to the relatively small size of the samples.[6]

We also analyzed the features and the POS that were predicted better, in comparison to Chersoni et al. (2021)'s experiment using English data (see Figure 1). Our analyses revealed that, similarly to English, the predictions for the COGNITION domain were the best. This is not surprising, because this domain is important for characterizing abstract concepts, of which textual/linguistic information is probably the prevailing source for human concept learning (Vigliocco et al., 2009). Sensory and Motor features were also predicted at relatively high correlations level, suggesting that many aspects of experiential, first-hand information can still be retrieved from linguistic data (Riordan and Jones, 2011). Finally, domains related to Spatial, Temporal (NUMBER and EVENT) and Attention turned out to be most challenging ones, coherently with the findings of Chersoni et al. (2021)'s experiment.

It can also be seen that, while English nouns were predicted much better than other POS, similar correlations were observed for nouns and verbs in Chinese (adjectives were the most difficult in both languages).

6 Conclusions

In this paper, we introduced Binder-style norms for Mandarin Chinese, collected using a similar method to the original study, and ran regression experiments from embeddings to norms, showing that the latter can be predicted with moderate to high correlations with humans. Such an application is especially interesting because it allows to extend the norms to large portions of the lexicon.

In the future, we plan to experiment with regression models based on contextualized vectors and to run tests for zero-shot crosslingual norms predictions, which could pave the way for the automatic acquisition of norms in low-resource languages.

[6] p-values computed with Fisher's r-to-z transformation.

Acknowledgments

This work has been supported by the Start-up Fund for new recruits of the Hong Kong Polytechnic University (1-BE8G) and the RGC Direct Allocation Grant (A-PB1C).

References

Samira Abnar, Rasyan Ahmed, Max Mijnheer, and Willem Zuidema. 2018. Experiential, Distributional and Dependency-Based Word Embeddings Have Complementary Roles in Decoding Brain Activity. In *Proceedings of the LSA Workshop on Cognitive Modeling and Computational Linguistics*.

Jeffrey R Binder, Lisa L Conant, Colin J Humphries, Leonardo Fernandino, Stephen B Simons, Mario Aguilar, and Rutvik H Desai. 2016. Toward a Brain-Based Componential Semantic Representation. *Cognitive Neuropsychology*, 33(3-4):130–174.

Damián E Blasi, Joseph Henrich, Evangelia Adamou, David Kemmerer, and Asifa Majid. 2022. Over-reliance on English Hinders Cognitive Science. *Trends in Cognitive Sciences*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Erin M Buchanan, Kathrene D Valentine, and Nicholas P Maxwell. 2019. English Semantic Feature Production Norms: An Extended Database of 4436 Concepts. *Behavior Research Methods*, 51:1849–1863.

Luana Bulat, Stephen Clark, and Ekaterina Shutova. 2017. Speaking, Seeing, Understanding: Correlating Semantic Models with Conceptual Representation in the Brain. In *Proceedings of EMNLP*.

Luana Bulat, Douwe Kiela, and Stephen Christopher Clark. 2016. Vision and Feature Norms: Improving Automatic Feature Norm Learning Through Cross-Modal Maps. In *Proceedings of NAACL-HLT*.

John A Bullinaria and Joseph P Levy. 2007. Extracting Semantic Representations from Word Co-Occurrence Statistics: A Computational Study. *Behavior Research Methods*, 39:510–526.

Emmanuele Chersoni, Enrico Santus, Chu-Ren Huang, and Alessandro Lenci. 2021. Decoding Word Embeddings with Brain-based Semantic Features. *Computational Linguistics*, 47(3):663–698.

Emmanuele Chersoni, Rong Xiang, Qin Lu, and Chu-Ren Huang. 2020. Automatic Learning of Modality Exclusivity Norms with Crosslingual Word Embeddings. In *Proceedings of *SEM*.

Kenneth Church and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1):22–29.

Steven Derby, Paul Miller, and Barry Devereux. 2019. Feature2Vec: Distributional Semantic Modelling of Human Property Knowledge. In *Proceedings of EMNLP*.

Barry Devereux, Colin Kelly, and Anna Korhonen. 2010. Using fMRI Activation to Conceptual Stimuli to Evaluate Methods for Extracting Conceptual Representations from Corpora. In *Proceedings of the NAACL Workshop on Computational Neurolinguistics*.

Barry J. Devereux, Lorraine K. Tyler, Jeroen Geertzen, and Billi Randall. 2014. The Centre for Speech, Language and the Brain (CSLB) Concept Property Norms. *Behavior Research Methods*, 46(4):1119–1127.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

Luana Fagarasan, Eva Maria Vecchi, and Stephen Clark. 2015. From Distributional Semantics to Feature Norms: Grounding Semantic Models in Human Perceptual Data. In *Proceedings of IWCS*.

Kimberly Glasgow, Matthew Roos, Amy Haufler, Mark Chevillet, and Michael Wolmetz. 2016. Evaluating Semantic Models with Word-Sentence Relatedness. *arXiv preprint arXiv:1603.07253*.

Alessandro Lenci, Magnus Sahlgren, Patrick Jeuniaux, Amaru Cuba Gyllensten, and Martina Miliani. 2022. A Comparative Evaluation and Analysis of Three Generations of Distributional Semantic Models. *Language Resources and Evaluation*, 56(4):1269–1313.

Ken McRae, George S. Cree, Mark S. Seidenberg, and Chris McNorgan. 2005. Semantic Feature Production Norms for a Large Set of Living and Nonliving Things. *Behavior Research Methods*, 37(4):547–559.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just. 2008. Predicting Human Brain Activity Associated with the Meanings of Nouns. *Science*, 320(5880):1191–1195.

Brian Murphy, Partha Talukdar, and Tom Mitchell. 2012. Selecting Corpus-Semantic Models for Neurolinguistic Decoding. In *Proceedings of *SEM*.

Gregory Murphy. 2002. *The Big Book of Concepts*. MIT Press, Cambridge, MA.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Francisco Pereira, Matthew Botvinick, and Greg Detre. 2013. Using Wikipedia to Learn Semantic Feature Representations of Concrete Concepts in Neuroimaging Experiments. *Artificial Intelligence*, 194:240–252.

Francisco Pereira, Greg Detre, and Matthew Botvinick. 2011. Generating Text from Functional Brain Images. *Frontiers in Human Neuroscience*, 5:72.

Yuanyuan Qiu, Hongzheng Li, Shen Li, Yingdi Jiang, Renfen Hu, and Lijiao Yang. 2018. Revisiting Correlations between Intrinsic and Extrinsic Evaluations of Word Embeddings. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 209–221. Springer.

Brian Riordan and Michael N. Jones. 2011. Redundancy in Perceptual and Linguistic Experience: Comparing Feature-based and Distributional Models of Semantic Representation. *Topics in Cognitive Science*, 3(2):303–345.

Jacob Turton, David Vinson, and Robert Smith. 2020. Extrapolating Binder Style Word Embeddings to New Words. In *Proceedings of the LREC Workshop on Linguistic and Neurocognitive Resources*.

Akira Utsumi. 2018. A Neurobiologically Motivated Analysis of Distributional Semantic Models. In *Proceedings of CogSci*.

Gabriella Vigliocco, Lotte Meteyard, Mark Andrews, and Stavroula Kousta. 2009. Toward a Theory of Semantic Representation. *Language and Cognition*, 1(2):219–247.

Gabriella Vigliocco and David P. Vinson. 2007. Semantic Representation. In Gareth Gaskell, editor, *The Oxford Handbook of Psycholinguistics*, pages 195–215. Oxford University Press, Oxford.

David P. Vinson and Gabriella Vigliocco. 2008. Semantic Feature Production Norms for a Large Set of Objects and Events. *Behavior Research Methods*, 40(1):183–190.

Shaonan Wang, Yunhao Zhang, Xiaohan Zhang, Jingyuan Sun, Nan Lin, Jiajun Zhang, and Chengqing Zong. 2022. An fMRI Dataset for Concept Representation with Semantic Feature Annotations. *Scientific Data*, 9(1):721.