Chinese and English speakers' neural representations of word meaning offer a different picture of cross-language semantics than corpus and behavioral measures

Benjamin D. Zinszer (bzinszer@ur.rochester.edu)
Andrew J. Anderson (andrewanderson@ur.rochester.edu)
Rajeev D. S. Raizada (rajeev.raizada@gmail.com)
Rochester Center for Brain Imaging
Department of Brain and Cognitive Sciences
University of Rochester
Rochester, NY 14627 USA

Abstract

Speakers of Chinese and English share decodable neural semantic representations, which can be elicited by words in each language. We explore various, common models of semantic representation and their correspondences to each other and to these neural representations. Despite very strong cross-language similarity in the neural data, we find that two versions of a corpus-based semantic model do not show the same strong correlation between languages. Behavior-based models better approximate cross-language similarity, but these models also fail to explain the similarities observed in the neural data. Although none of the examined models explain cross-language neural similarity, we explore how they might provide additional information over and above cross-language neural similarity. We find that native speakers' ratings of noun-noun similarity and one of the corpus models do further correlate with neural data after accounting for cross-language similarities.

Keywords: cross-language semantics, multivoxel pattern analysis, semantic models, concept representation

Introduction

Multi-voxel pattern analyses and neural decoding methods have recently enabled cognitive neuroscientists to predict patterns of brain activity for word stimuli by generalizing from a training set of words and their associated functional neuroimaging data to new words (e.g., Mitchell et al., 2008) or new participants (e.g., Raizada & Connolly, 2012). One application of this inference has been using functional brain data to link words across languages, known as neural translation. Studies of bilingual speakers have shown decodable associations between a speaker's mental representations of the same word in each of their two languages (Buchweitz et al., 2012; Correia et al., 2014). In recent work, we translated a small lexicon of seven words between native speakers for Chinese and English by comparing each group's functional brain activity in representational similarity space (Zinszer et al., 2015).

Findings in neural translation are generally consistent with hypotheses that suggest semantic representations should be strongly correlated across speakers based on shared extrinsic (e.g., sensory) information (Binder & Desai, 2011; Hauk et al., 2006). For example, the appearance, sound, and general functions of a dog would be roughly the same for speakers of any language. Zinszer et al.'s (2015) neurally-based translation provides evidence that word representations do encode this perceptual information. Thus, on one hand, neural representations of translation equivalent words should be the same to the extent that experience in the world is shared across speakers of different languages. On the other hand, research in translation ambiguity shows that even translation equivalents evoke unique semantic information across languages (Degani & Tokowicz, 2010; Malt & Majid, 2013). Highly accurate models of word meaning should reflect some language-specific information and diverge across languages.

In corpus-based models, word co-occurrence statistics differ between translation-equivalent words when those words are used in systematically different ways across languages. This problem is widely known in cross-language information retrieval (CLIR) and has inspired a number of approaches for improving translation accuracy (Zhou et al., 2012). However, many translation models rely on parallel documents and contextual clues to compare translations for a particular token. This information is not available to human participants or models for the isolated single-word stimuli used in most neural decoding studies.

Corpus-based semantic models reflect something important about word meaning as it is represented in the brain, as evidenced by their successful application to neural decoding (as in Mitchell et al., 2008), but the cross-language semantic differences that have vexed machine translation do not seem to prevent neural translation of small lexicons (Zinszer et al., 2015). In this sense, corpus-based models of word meaning may over-estimate these differences relative to speakers' actual mental representations.

In this study, we contrast three classes of model for word meaning: (1) Neural data for seven translation-equivalent concrete nouns in Chinese and English obtained by Zinszer et al. (2015), (2) two simple corpus-based models for the seven words in each language, and (3) behavioral models of word representation elicited from native speakers of each language. We compare these models using representational similarity analysis, evaluating cross-language similarity within the models and comparing the corpus- and behavior-based models to the neural data.

We predicted that corpus- and behavior-based models of word representation would reflect the translation relationships between words through high cross-language

correlations, although we expected this correlation to be somewhat attenuated in the corpus models given the nuances of translation ambiguity captured in the source material. We expect these models to be strongly correlated to the neural data. However, if the corpus- or behavior-based models also capture language-specific aspects of meaning, we predict that the models will correlate with variance in neural data *not* explained by the cross-language similarity between the neural data of the two groups.

Method

Target Nouns

The target nouns were seven translation-equivalent words in English and Chinese. These words were selected based on four requirements: (1) concrete nouns, (2) monosyllabic in both languages, (3) represented by a single Chinese character, and (4) unlikely for English translations to be known by the Chinese participants (see Table 1 for list). To verify criterion (4), Chinese participants in the fMRI study completed a brief quiz in which they wrote the English translations for twenty Chinese words. Mean translation accuracy was 1.56 of the 7 target nouns. In the following analyses, we describe a variety of different multivariate representations for these target nouns based on data from neuroimaging, corpus, and behavioral measures.

Table 1. Target nouns (first and bolded) and semantic associates in English and Chinese (with pinyin)

English	Chinese
axe, knife, sword	**斧 (fǔ)**，锯，剑
broom, mop, vacuum	**帚 (zhǒu)**，拖把，簸箕
gown, dress, robe	**袍 (páo)**，裙，礼服
hoof, paw, foot	**蹄 (tí)**，爪，脚
jaw, mouth, bone	**颚 (è)**，口，骨
mule, horse, donkey	**骡 (luó)**，马，驴
raft, boat, barge	**筏 (fá)**，船，艇

Neural Representations

Functional MRI data were acquired from Zinszer et al.'s (2015) neural translation experiment. In that experiment, eleven native speakers of English (4 M / 7 F) and Mandarin Chinese (3 M / 8 F) at Dartmouth College completed a simple semantic relatedness task while undergoing fMRI. Participants viewed 49 words (7 target nouns, 42 filler) in pseudorandom order, repeated over seven runs. To ensure semantic processing of the stimulus words, they were periodically asked to rate the semantic relatedness of a word to the preceding word. Individual participants' neural responses were estimated for the seven target nouns based on a GLM model of functional activity. Each participant's functional responses were abstracted into similarity space (see "Representational Similarity Analysis") and these similarity structures were averaged across participants within the same language group. Similarity structures were estimated for whole brain data as well as for each of 96 ROIs in the Harvard-Oxford neuroanatomical atlas. Further details of these procedures are described in the original study (Zinszer et al., 2015).

Corpus-Based Representations

Following the method of Mitchell et al. (2008), semantic representations were constructed for each target noun based on their co-occurrence rates with 25 verbs in a large corpus of webpages. The 25 Chinese verbs were obtained from five Chinese-English bilinguals who translated the English verbs independently and were then aggregated based on majority consensus (see Table 2). In a few cases, two English verbs most frequently translated to the same Chinese verb (e.g., hear and listen as 听). In these cases, a lower frequency or compound Chinese verb with similar meaning was substituted to maintain 25 unique dimensions.

We used the Leeds University query tool for two similar Internet corpora in Chinese (90 million words) and English (160 million words; http://corpus.leeds.ac.uk/internet.html; Sharoff, 2006) to obtain concordance counts in a three word window for each noun-verb pair. Count data vectors were normed to unit length, yielding a 25-dimensional vector for each noun.

Due to the low frequency of some target nouns, we also generated broadened representations from the same corpus data. For each target noun, two associated nouns were selected to help bring out defining elements of the target noun relative to noise-level co-occurrences (e.g., *mop* and *vacuum* capture information about shape and function of *broom*). The three 25-d vectors for the target noun and its two associates were averaged to produce the broadened representation. Table 1 lists associates of each target noun.

Table 2. Verbs in English and Chinese

English	Chinese	English	Chinese
see	看	enter	进入
say	说	move	移动
taste	尝	listen	倾听
wear	穿	approach	接近
open	开	fill	填
run	跑	clean	清理
near	靠近	lift	举
eat	吃	rub	擦
hear	听	smell	闻
drive	驾驶	fear	怕
ride	骑	push	推
touch	碰	manipulate	操纵
break	打破		

Behavior-Based Representations

Eleven native English speakers (4 M / 7 F) and eleven native Chinese speakers (6 M / 5 F) at the University of Rochester (Rochester, NY, USA) completed two semantic

relatedness judgment tasks in their respective native languages.

In the first task (hereafter the NN ratings), participants judged the semantic relatedness for every pairwise combination of the seven target nouns (21 total comparisons). Order of words in each pair and order of pairs were randomized for each participant. Semantic relatedness judgments were made on a scale of 1 (unrelated) to 5 (highly related) and averaged over participants.

After an unrelated intervening task (visual categorization of animal-like figures), participants made binary semantic relatedness judgments for every target noun with each of the 25 verbs (175 total binary ratings, hereafter the NV ratings). Responses were averaged across participants to yield decimal values for each NV pair.

Representational Similarity Analysis

Representational similarity analysis compares the way a set of referents (such as the seven target nouns) is organized in different representational spaces. For example, multivariate brain activation patterns and corpus-based models described in the preceding sections can be compared to each other when each is abstracted into similarity space. This similarity space is composed of Pearson pairwise correlations between the multivariate representations for each of the seven target nouns, resulting in a 7x7 similarity matrix in which each noun is described by the correlation of its multivariate representation to those of the other six nouns. The correlation values are further transformed using Fisher's *r*-to-*z* for normalizing correlation coefficients. We can then compare two 7x7 similarity spaces by Pearson correlation of their 21 unique values (the lower-left triangle of the matrix).

Results

We computed similarity structures for the each measure of neural and semantic relatedness. The NN ratings were used directly by averaging the English and Chinese participants' responses to the semantic relatedness task for each noun .

Cross-Language Correlations

For all measures, 21 unique values from the 7x7 similarity matrices (as in Figure 1) were correlated across languages, such as NN English vs. NN Chinese. Table 3 reports these cross-language correlations. The whole brain fMRI data and both sets of behavioral ratings (NN and NV) reflected strong cross-language correlations (r<0.60, p<=0.003). The Leeds corpus measure showed no such cross-language correlation, although the Broadened Leeds representations were moderately correlated (r=0.44, p<0.05).

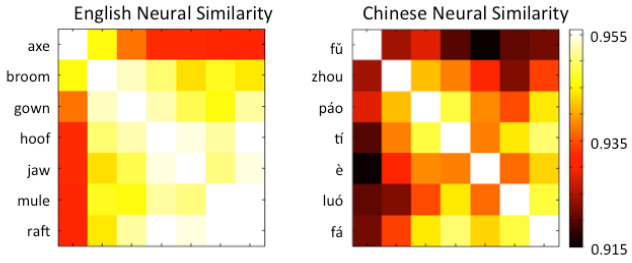

Figure 1. Neural similarity in each language group. Comparable similarity matrices were computed for the target nouns in each corpus and behavioral model, allowing cross-language and cross-model comparisons.

Table 3. Cross-language correlations for each measure of word representation

Measure	*r*	*p*
MRI - Whole brain	0.89	< 0.001
NN rating	0.85	< 0.001
NV rating	0.61	0.003
Broadened Leeds	0.44	0.045
Leeds	-0.08	0.726

Model-to-Brain Correlations

Next we compared each semantic model to fMRI data from the corresponding language to measure the degree to which patterns of functional activity correspond to the words' semantic representations in the models. These correlations are reported in Table 4 for the whole brain data, and summary statistics are provided for correlations across 96 anatomical ROIs as defined by the Harvard-Oxford brain atlas (http://www.fmrib.ox.ac.uk/fsl/).

Table 4. Correlation of each semantic model to whole brain neural data and summary statistics for correlations at ROIs.

	Whole brain		Harv.-Oxf. ROIs		
Model	***r***	***p***	**mean *r***	***s.d.***	**max |*r*|**
English					
NN	0.15	0.52	0.16	0.13	0.47
NV	0.11	0.64	0.05	0.13	0.49
Broad	-0.24	0.30	-0.14	0.17	0.60
Leeds	-0.24	0.30	-0.13	0.15	0.41
Chinese					
NN	-0.11	0.62	-0.09	0.13	-0.36
NV	-0.08	0.72	-0.10	0.14	-0.43
Broad	-0.47	0.03	-0.45	0.11	-0.69
Leeds	-0.17	0.46	-0.15	0.10	-0.38

All semantic models and the two sets of whole brain data were plotted using multi-dimensional scaling for cosine similarity (Figure 2, next page). As suggested by the correlations to brain data reported in Table 4, Chinese and English speakers' patterns of functional response were closer to one another than they were to any of the semantic models. Behavior-based models (NN and NV) clustered together in the MDS projection, and the NN models were both the most similar across languages and the closest approximations of the brain data. The Leeds corpus-based representations were distantly separated from one another and from the brain data, but the Broadened model in

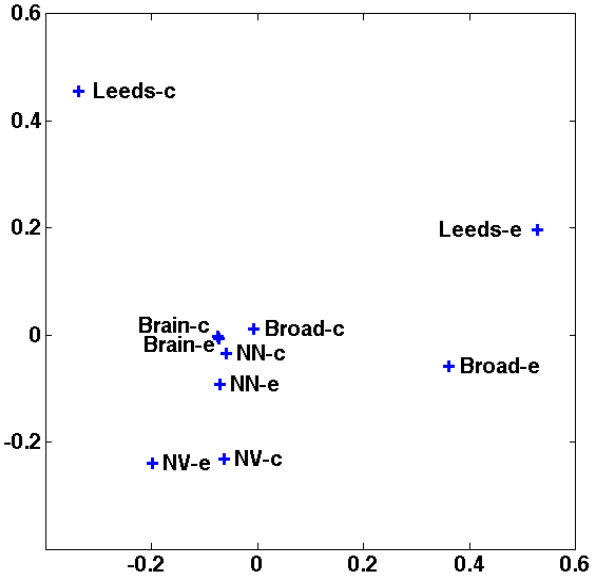

Figure 2. MDS plot of similarity space for semantic models and whole brain representations in Chinese (c) and English (e). Brain-e and Brain-c representations are overlapping.

Chinese more strongly (though negatively) correlated with the brain data.

Residuals Analysis

The four semantic models offered a poor account of the neural responses in both languages, suggesting that the cross-language similarity supporting neurally-based translation was not explained by any of the proposed corpus- or behavior-based models. In our next analysis, we turned our attention to the language-specific components of the neural representations by isolating the residuals of the neural representations in each language that remained after accounting for cross-language neural correlations.

In this residuals analysis, we again correlated the semantic models (behavior- and corpus-based) to neural data, but we evaluated the models against only the unexplained variance in the Chinese and English neural data that remained after cross-language correlation (that is, the residuals of the correlation). Here we ask to what degree our semantic models explain the remaining variance between the Chinese and English neural representations after accounting for the language-invariant components.

For each ROI, we saved the residuals from the correlation between the Chinese and English neural representations and used them in lieu of the original neural representations. We then repeated the correlations reported in the preceding section. We found that across the 96 ROIs, correlations were generally weak and distributed around zero (see Table 5).

Because this analysis is principally concerned with the language-specific contribution of the semantic models after controlling for cross-language similarity, we searched for ROIs where the *both* correlation between the English model and English residuals *and* the correlation between the Chinese model and Chinese residuals were significant. Thus, each ROI has a p-value indicating the probability of both languages' models producing significant correlations under the null hypothesis. Using this constraint, a critical p-value of $\sqrt{\alpha}$ in *each* language produces a joint probability of Type I error equal to α for an ROI if the null hypothesis is true. Bonferroni correction for multiple comparisons across the 96 ROIs yields a threshold of $\alpha = 0.00052$ for each ROI, and thus we highlighted ROIs where *both* the Chinese and English correlations were $p < \sqrt{0.00052}$ or 0.023.

Table 5. Summary statistics for correlations of each semantic model to neural data in Harvard-Oxford ROIs.

	Harv.-Oxf. ROIs		
Model	**mean *r***	***s.d.***	**max |*r*|**
English			
NN	0.14	0.18	0.56
NV	-0.02	0.20	0.42
Broad	-0.03	0.23	0.63
Leeds	0.07	0.19	0.40
Chinese			
NN	-0.11	0.19	-0.66
NV	-0.08	0.21	-0.55
Broad	-0.30	0.18	-0.69
Leeds	-0.10	0.15	-0.39

Table 6. Cortical regions in which a semantic model significantly correlates with neural representations in the same language after controlling for cross-language neural similarity.

			Model-to-brain	
Model	**Harvard-Oxford ROI**		**Eng. *r***	**Chi. *r***
NN	3	R Middle Frontal Gyrus	0.52	-0.53
	16	R Postcentral Gyrus	0.56	-0.66
Broad	40	R Frontal Operculum	0.64	-0.53
	46	L Supracalcarine Gyrus	0.51	-0.61

The NN behavioral model and Broadened Leeds corpus model yielded significant results in both languages. Table 6 lists the regions where English and Chinese models explained additional variance in their respective languages beyond the cross-language neural similarity. We also investigated whether the semantic models in each language exclusively correlated to the neural data for that language, or whether they also correlated with the opposite language. This test of double-dissociation identifies whether the semantic model is using language-specific information to explain variance in the target language (e.g., English model only correlates with English neural data) or language-invariant information (e.g., English model correlates with both English and Chinese neural data). In all four regions of interest identified in Table 6 we found that at least one of the models (English or Chinese) significantly correlated with the neural data in both languages (p<0.05).

Discussion

In this paper, we compared four semantic models (two corpus-based and two behavior-based) in Chinese and English with neural similarity structures that have previously been used to translate between speakers of English and Chinese. The strong cross-language correlations in the neural data reported by Zinszer et al. (2015) indicate that important language-invariant information underlies the neural responses to word stimuli. Cross-language correlations revealed in the behavioral models also supported this conclusion, but these models did not correlate with the neural responses, pointing to another source of information. Given the previous success of using corpus-based models of word meaning to decode neural data (e.g., Mitchell et al., 2008), we reasoned that language-specific aspects of word meaning might be captured by our models and offer insight into the what aspects of the neural representations were language-specific and what aspects were attributable to universally available conceptual information (i.e., world-knowledge).

Representations Shared Across Languages

The strong cross-language correlations in the neural data indicate that language-invariant semantic information underlies a significant portion of the neural responses to word stimuli. Although neither the corpus nor behavioral models showed the same degree of cross-language convergence, the NN ratings were very close despite the potential differences in word meanings across languages (Degani & Tokowicz, 2010). In this sense, directly querying native speakers' intuitions about the similarity in word meanings (the NN model) produced the most accurate reflection of cross-language similarities in the neural representation of word meaning. Analogous ratings of object similarity have proved useful in identifying individual differences in the neural representations of objects (Charest et al., 2014), so this close link between direct behavioral query of a representation (the NN ratings) and neural similarity is not surprising.

Surprisingly, correlations between the whole-brain data and all four semantic models proved almost entirely insignificant, with the exception of the Broadened Leeds model's correlation with the Chinese neural data. Searching the 96 cortical regions of the Harvard-Oxford atlas did not yield strong correlations between the neural representations and most of the semantic models. Only the Broadened Leeds corpus showed an overall trend towards correlating with the Chinese data, but none of these correlations survived correction for multiple comparisons.

Taken as a whole, these results are suggestive of at least two important underlying sources of information. The regularities in similarity structures between languages suggest that both the neural data and behavioral model can only be explained by language-invariant representations, and the failure of these two measures to correlate with one another suggests that they may capture different aspects of semantic knowledge. The corpus-based models did not show as much cross-language regularity, nor did they correlate with the neural data, leaving their meaning ambiguous in this context.

Language-Specific Representations

The initial corpus-based semantic models were extremely divergent from one another across languages, suggesting that they may encode highly language-specific aspects of word meaning. These models perhaps even exaggerate the differences between speakers' semantic representations in Chinese and English, in light of the high correlations across the behavioral models. However, this divergence is not as surprising when one considers the broad structural differences between written Chinese and English.

Although neither the corpus nor behavioral models showed explanatory power for the neural representations, we show in the residuals analysis that some divergent information is contained in this neural signal. The noun-noun similarity ratings and Broadened Leeds corpus models both added explanatory power over and above the direct comparison of neural representations in each language, confirming that these models encode real, neurally-implemented information that is not shared across the speakers' neural responses in each language. However, this explanatory power was (paradoxically) not language-specific. Every region in which the models significantly correlated with their own language was also explained by at least one model from the other language. This observation challenges the assumption that the NN or Broadened Leeds semantic models represent purely language-specific information. An alternate explanation may be that language-invariant information encoded in these models is differentially represented in the brains of speakers of one language or the other.

The ROIs showing the strongest such relationships are also intriguing. Bilateral postcentral gyri produced strong correlations between languages in Zinszer et al.'s (2015) cross-language comparison of neural data, a fact that was taken as evidence for language-independent somatosensory involvement. The correlation between representations in the L Supracalcarine gyrus and the Broadened Leeds corpus are somewhat puzzling, given this region's importance in visual processing. The corpus model is not particularly visually oriented, and one would not expect visual correlates of meaning to be more apparent in a model of word co-occurrence than in a behavioral model or cross-language neural similarity.

Future Work

Several perplexing and tantalizing questions arise from these comparisons of neural, corpus, and behavioral models of word meaning in Chinese and English. One major limitation of the present study is the amount of statistical power available from an analysis of only seven words in each language. While many of our investigations yielded null results, this finding could arise from a genuinely null

relationship or from the impoverished representation of each language in this highly constrained set of stimuli.

Further, the meaning of significant negative correlation between the broadened Chinese corpus model and Chinese neural data is not immediately clear. This correlation indicates regularity in both structures (less similarity in the corpus reliably corresponds to greater similarity in the brain), but we propose that these measures may be mutually informative in future studies of neural decoding. Our own informal investigations suggest that the co-occurrence relationships of interest may be more remote than detectable in the three word window, thus skewing the present model based on syntactic constraints.

Finally, while we know that functional responses of Chinese and English speakers contain information that can discriminate between words in both languages, this shared information is still not adequately explained by any individual model described in this paper. What then do each of the neural, behavioral, and corpus models represent? Recent neural decoding research has focused on decomposing the neural signal into interpretable components of meaning (e.g., sensory modalities or cognitively plausible features). An integrative approach has indeed proved important for comparing experiential and corpus-based representations (Andrews et al., 2009). Thus evaluating a more elaborated set of stimulus words with a combination of these more detailed models may be instrumental in unlocking these various sources of semantic information encoded in the brain.

Conclusions

We know that cross-language neural similarity is meaningful because it permits decoding across languages, but whatever the source of these cross-speaker and cross-language regularities, it is not directly derived from speakers' intuitions about semantic relatedness nor from corpus statistics, nor are these sources of information irrelevant to neural representation of semantics since they do provide some explanatory power beyond the cross-language correlations. Alternate representational models and more representative stimulus words may yet provide better descriptions of cross-language semantics and their respective implementations in the brains of a language's speakers. Future research will test such models' ability to decode word-elicited concepts across languages and clarify this highly complex emerging picture.

Acknowledgements

We are grateful to Donias Doko and Carol Jew for data collection, Xixi Wang for language consulting, and Peiyao Chen, Anqi Li, Yinghui Qiu, and Tianyang Zhang for translation.

References

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review, 116*(3), 463-493.

Binder, J. R., & Desai, R. H. R. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences, 15*(11), 527–36. doi:10.1016/j.tics.2011.10.001

Buchweitz, A., Shinkareva, S. V, Mason, R., Mitchell, T. M., & Just, M. A. (2012). Identifying bilingual semantic neural representations across languages. *Brain and Language*, *120*(3), 282-9. doi:10.1016/j.bandl.2011.09.003

Charest, I., Kievit, R. a., Schmitz, T. W., Deca, D., & Kriegeskorte, N. (2014). Unique semantic space in the brain of each beholder predicts perceived similarity. *Proceedings of the National Academy of Sciences*, *111*(40), 14565–14570. doi:10.1073/pnas.1402594111

Correia, J., Formisano, E., Valente, G., Hausfeld, L., Jansma, B., & Bonte, M. (2014). Brain-based translation: fMRI decoding of spoken words in bilinguals reveals language-independent semantic representations in anterior temporal lobe. *The Journal of Neuroscience*, *34*(1), 332–8. doi:10.1523/JNEUROSCI.1302-13.2014

Degani, T., & Tokowicz, N. (2010). Semantic ambiguity within and across languages: an integrative review. *Quarterly Journal of Experimental Psychology, 63*(7), 1266–303. doi:10.1080/17470210903377372.

Hauk, O., Johnsrude, I., & Pulvermu, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, *41*, 301–307.

Malt, B. C., & Majid, A. (2013). How thought is mapped into words. *Wiley Interdisciplinary Reviews: Cognitive Science*. doi:10.1002/wcs.1251

Mitchell, T. M., Shinkareva, S. V, Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A, & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science (New York, N.Y.)*, *320*(5880), 1191–5. doi:10.1126/science.1152876

Raizada, R., & Connolly, A. (2012). What makes different people's representations alike: Neural similarity space solves the problem of across-subject fMRI decoding. *Journal of Cognitive Neuroscience*, *24*(4), 868–877.

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. In Marco Baroni and Silvia Bernardini, (eds), *WaCky! Working papers on the Web as Corpus.* Gedit, Bologna, http://wackybook.sslmit.unibo.it/

Zhou, D., Truran, M., Brailsford, T., Wade, V., & Ashman, H. (2012). Translation techniques in cross-language information retrieval. *ACM Computing Surveys*, *45*(1), 1–44. doi:10.1145/2379776.2379777

Zinszer, B. D., Anderson, A. J., Kang, O., Wheatley, T., & Raizada, R. (2015). You say potato, I say tudou: How speakers of different languages share the same concept. *Proceedings of the 37th annual conference of the Cognitive Science Society.*