Commentary on Kemmerer: The challenges and rewards of trying to combine linguistics and cognitive neuroscience

Elizabeth A. Shay [1], Scott Grimm [2] and Rajeev D. S. Raizada [1]

[1]Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY, 14627

[2]Department of Linguistics, University of Rochester, Rochester, NY, 14627

Introduction

Kemmerer's target article points to the necessity of conducting cognitive neuroscience research cross-linguistically, particularly considering languages with nominal classification systems (Kemmerer, 2016). The particulars of focusing on this type of research points to a larger idea that needs emphasis: linguistics and cognitive neuroscience need to work together more in interdisciplinary teams to understand language, and its connection to the brain, fully.

This type of collaboration, supported by Kemmerer (2016), has the potential to provide fruitful avenues for language research. We are currently working in one such collaboration, and there are several difficulties that can arise, likely leading to the current fragmentation of these two areas of language science. In the following sections, we will discuss some of these challenges in collaborating, ways to connect these two fields, and some tractable problems in which this methodology allows for exploration.

Difficulties in collaboration between linguistics and cognitive neuroscience

Poeppel and Embick (2005) described two overarching problems of collaborating between linguistics and cognitive neuroscience. The first of these is the problem that linguistics and neuroscience work at different levels of granularity. More specifically, they mention that linguistics has many fine-grained distinctions and explicit computations. Cognitive neuroscience, on the other hand, deals with broader conceptual differences. The other challenge the researchers focus on is the ontological incommensurability problem: the units used in the two fields are not the same. For instance, they describe

linguistics as working in distinctive features, syllables, morphemes, noun phrases, and clauses, whereas neuroscience works in dendrites and spines, neurons, cell-assembly or ensembles, populations, and cortical columns. Determining the mapping between these two sets of units is difficult. We add to this that the two fields also differ as to the units of language that they examine. Linguistics can much more easily, and frequently does, look at full utterances or even large spans of discourse. Cognitive neuroscience is just now starting to look at these questions and primarily focusing on single words or perhaps short phrases.

Another challenge within collaborations between cognitive neuroscience and linguistics is the difference in focus on aspects of the same problem. Both fields are interested in questions related to semantic representations, but this typically means different things to researchers from the two fields. Cognitive neuroscience often equates semantics with determining the contents of a concept in terms of activated brain regions. For instance, the popular embodied cognition framework (e.g. Pulvermüller and Fadiga, 2010) examines the perceptual and sensory-motor regions that are activated when thinking about the meaning of certain words. In linguistics, by contrast, semantic representations, at least in the dominant tradition of truth-conditional model-theoretic semantics, are founded upon referents, viz. actual things in the world, and the combination of meanings in terms of referents which form sentences which can be evaluated as either true or false (e.g. the work of Frege, Montague, and others. See Dowty et al. (1981) for a classic textbook introduction).

In neither tradition is the relation between referent and concept well-articulated: model-theoretic semantics focuses on representations based on referents at the expense of representing concepts, whereas cognitive neuroscience tends to have a bias in the opposite direction. Thus, while both fields legitimately investigate "meaning", the meaning of "meaning" differs in the two fields, one of many terminological challenges facing interdisciplinary teams of linguistics and cognitive neuroscience that must be navigated. This is a tension that appears in Kemmerer (2016), which focuses on the relevance of classifiers for objects concepts, in contrast to most of the linguistics literature which focuses on classifiers, in terms of their semantics, as applying to referents or, as a formal grammatical device, applying to nouns, but not to object concepts *per se* (see discussion in Contini-Morava and Kilarski (2013)).

A final difference we will discuss here between linguistics and cognitive neuroscience is that the two fields work at two different Marr

levels (Marr, 1982). Linguistics focuses on the computational level, considering possible ways we could "solve" language, with little or no consideration about how this might actually work in humans. Cognitive neuroscience works at the implementational level, considering how the brain deals with small chunks of language with, too often, little or no consideration about how this might fit into the larger language picture across all of a single language, let alone cross-linguistically.

Bridging the gap: studying representations instead of activations

We wish to argue here that the core of this granularity problem has been the question of studying representations. Until recently, human neuroimaging studies were restricted solely to measuring activation, and could only make broad-brush statements about the overall sort of information that was being represented. For example, studies might make statements of the sort "The ventral temporal cortex is involved in representing semantic information about objects". More recently, however, fMRI studies have been able to probe much finer-grained aspects of neural representational structure, e.g. measuring the pattern similarities between the neural representations of specific individual words (Mitchell et al., 2008; Anderson et al., 2016b) or, more recently, full sentences (Anderson et al., 2016a). By studying representations of linguistic meaning, cognitive neuroscience can now engage much more directly with linguistics, as both disciplines are now at last able to study the same subject matter.

The previous inability of cognitive neuroscience to address this representational level was a contributing factor to differences in the way the two fields would talk about a given topic. For example, when papers in cognitive neuroscience referred to "semantic processing", they typically meant "the brain areas that are active during a semantic task", whereas when linguistics papers talked about semantic processing they typically meant "semantic representations and the information processing operations that are applied to them."

The key bridge between semantics and cognitive neuroscience has been to relate structure in a semantic space to structure in a space of multivoxel neural activation patterns. This work was initiated in the seminal paper by Mitchell et al. (2008), who used linear regression to learn a mapping between a vector space model of meaning and fMRI activation patterns.

In such work, the semantic information that is being related to brain activation patterns is that of vector space semantics (e.g. Turney and

Pantel, 2010), in which word meanings are represented as vectors of numbers. In corpus-based distributional semantics, those numbers are statistics, typically word co-occurrence frequencies, extracted from large bodies of text. These approaches instantiate the well-known saying of Firth (1957) "You shall know a word by the company it keeps". A different approach is behavioural (e.g., McRae et al., 2005; Binder et al., 2016), and is often known as feature norming or sometimes as experiential semantics: people are asked to make rating judgments about the degrees to which individual words have various featural properties, such as size, shape, being man-made and so on. The word meanings are then represented in terms of those behavioural features.

A crucial characteristic of a semantic vector space is that it has structure. Some meanings are closer together in semantic space, i.e. they have similar meanings, whereas others are further apart. For example, such a model would capture the fact that the word 'apple' is closer in meaning to 'pear' than it is to 'truck'. The locations of points in a space can be represented in many different coordinate systems. One powerful framework is that of a similarity space, i.e. representing items in terms of their similarities to each other, which has proven to be a powerful modeling tool across multiple areas of cognitive science and machine learning.

Just as meanings can be considered as situated in semantic similarity space, multivoxel fMRI activation patterns can be viewed as forming a neural similarity space. Individual fMRI voxel activations can only go up and down, but the activation across multiple voxels forms patterns, and those patterns can have varying degrees of similarity to each other. Moreover, mappings can be learned between one set of vectors and another, in this case between vectors representing word meanings in a semantic model, and vectors of multivoxel brain activation patterns (e.g. Mitchell et al., 2008).

Thus, recent work in the cognitive neuroscience of language, including work from our own group, has taken the approach of using vector space semantic models to decode vectors spaces of multivoxel neural activation patterns. This constitutes a substantive and new bridge between linguistics and cognitive neuroscience, but it still has many current limitations. For example, it interfaces with only one approach to semantics, namely the vector space approach. That approach has been popular in computational linguistics, but it does not (as yet) have a clear relation to the tradition in the formal semantics branch of linguistics of truth-conditional model-theoretic semantics (e.g. Dowty et al., 1981) (although see Baroni, Bernardi, and

Zamparelli (2014) among others for initial attempts at combining formal and distributional semantics). Indeed, representing linguistic sentential structure is a challenging problem for vector space semantics and for cognitive neuroscience work based upon it, although a few recent studies have started to make initial forays in this direction. For example, Frankland and Greene (2015) found neural activation patterns that were sensitive to agent-patient relationships in sentences, and in our own work (Anderson et al., 2016a) we were able to perform neural decoding of entire sentences. However, that sentence-level decoding is just an initial step with much future work to do, as it uses a "bag of words" modeling approach that is insensitive to word order within a sentence.

Approaches to relating linguistics to cognitive neuroscience: classifiers and beyond

Kemmerer's (2016) target article illustrates one way in which linguistics and cognitive neuroscience could conduct interdisciplinary research. Using the nominal classification system details of any particular language, as described by linguists, cognitive neuroscientists could get a more complete picture of the cross-linguistic variation in regions that activate for animacy, size, and other important aspects of nouns.

A second question Kemmerer (2016) considers is how "two explicitly coded levels of object categorization", a noun and its classifier, are "coordinated", in more traditional linguistics terminology how the meanings compose. This is a much more difficult problem than might first appear. The framing of the problem limits nominal classifiers to referencing superordinate information, but the use and meaning of classifiers within a language is far more complex. In fact, it has long been recognized that from a strictly semantic viewpoint the superordinate information referenced by classifiers is non-informative since, e.g., the concept *horse* entails *animal*, thus to have a classifier specifying that the horse is animate is redundant. Instead, research on classifiers has adduced evidence that classifiers often serve on the one hand as a derivation device and/or as part of a language's reference tracking systems, e.g. performing tasks similar to what definite or indefinite determiners (*the*, *a*) do in English (Aikhenvald (2000); Contini-Morava and Kilarski (2013)). Neither function is straightforward to examine in terms of how meanings compose. Derivational uses are parade cases of non-compositional behavior, while when used as reference tracking device, then understanding how

a noun and its classifier are coordinated may involve understanding how the referent is integrated into the sentence or discourse, and as pointed out, cognitive neuroscience approaches to meaning are just beginning to venture to exploring meaning on the sentential level.

Some interesting recent behavioural work (Speed et al., 2016) suggests that, although a language's use of classifiers may indeed be related to the conceptual structure of that language's speakers, the direction of causality may flow from from conceptual structure to classifier systems, rather than the other way around. Those authors asked people to rate the similarity of a variety of objects, with those people either being native speakers of a language that does *not* use classifiers, namely Dutch, or native speakers of a language that *does* use classifiers, namely Chinese. The experimenters asked their participants to compare some sample objects to a target object, and they designed their experiment such that one of the sample objects shared the same classifier (in Chinese) with the target, whereas the others did not. If the presence of this classifier were to influence judgments of similarity, then the Chinese speakers should have rated that classifier-sharing object as more similar to the target than the Dutch speakers did. However, no difference between the two language groups was observed. As the authors write: "This suggests that classifier systems reflect, rather than affect, conceptual structure."

We close by suggesting some tractable questions linguists and cognitive neuroscientists may focus on instead of or in addition to the questions raised by Kemmerer (2016).

Despite the difficulties examining how classifiers compose with nominals, we believe one tractable area for connection between linguistics and cognitive neuroscience is in semantic composition, in particular Shay and Raizada (2015) have done some work looking at basic composition functions over sensory-motor features. These sensory-motor features were developed from a meta-analysis of cognitive neuroscience studies and an additional neuroimaging study, suggesting that these particular attributes are highly associated with particular regions of the brain (Fernandino et al., 2015).

They found that addition of these sensory-motor feature vectors did remarkably well at predicting the correct phrase sensory-motor vector (Shay and Raizada, 2015). Ongoing research is connecting these findings with fMRI research on these same individual words and phrases. Based on the vector research and the neurobiological basis of these features, it should be expected that adding together the brain representations could make a reasonable approximation of the phrase activation patterns.

Thus, this ongoing line of our research allows us to relate semantic composition functions from linguistics to functions that are hypothesized to be acting upon the corresponding multivoxel neural activation patterns. Furthermore, the sensory-motor feature vectors can be used to probe other composition functions that are used within linguistics to get a more complete picture of this process.

A very valuable and central aspect of Kemmerer (2016) is that it highlights the importance of carrying out cross-linguistic studies in the cognitive neuroscience of language. Such studies may seek, as Kemmerer suggests, to reveal how differences between languages may be reflected in differences between the neural representations of those languages' speakers. Another approach, which we recently followed in Zinszer et al. (2016), is to examine semantic commonalities across languages which differ greatly in their surface characteristics. In that paper, we investigated Chinese and English: native Chinese-speakers were presented with words in Chinese, and native English speakers read the corresponding English-translation words (a set of seven concrete nouns). The only content in common across the two language groups was purely semantic, as the two languages are as orthographically and phonologically different as can be: Chinese words neither look, nor sound, at all like their English counterparts. We found that by matching the neural similarity structure of elicited brain activation across the two groups we were able to deduce the corresponding semantic matches. In other words, we could translate between English and Chinese words using only neural activation as our guide. Thus, the semantic structure that was shared across the languages was reflected in a shared neural representational structure across those languages' speakers.

Given these promising results in comparing semantic representations of nouns across languages, we see great potential to generalize this technique to address questions relating to semantic typology in the future. In particular, focusing on semantic representations brings us one step closer to being able to match the methodology of semantic typology in the fMRI lab. The central methodological tool in semantic typology is the 'etic grid': a constructed space of possible values for a semantic domain established along one or more different dimensions. In the literature on color categorization, different points in the space differ in their values along the dimensions of hue and brightness. In elicitation sessions, speakers are presented with a set of Munsell chips which code points in the color space. By naming each color chip with a color term, speakers establish

the reference of a color category in a language with respect to a set of hues (Berlin and Kay, 1991).

In the study of spatial semantic categories, such as those referenced by prepositions, the stimuli set consists of a set of pictures demonstrating spatial relations objects may stand in, such as an apple resting on a table (Bowerman, 1996), where the points on the etic grid are pictures of particular spatial scenes which vary along dimensions of spatial relation types, e.g. *in* or *under*. A common technique in semantic typology is to take this etic grad as the basis for a similarity space in which different language's semantic categories can be plotted and compared through multidimensional scale or other modeling techniques (Levinson et al., 2003; Bohnemeyer and Stolz, 2006). These methodologies cannot be implemented in the fMRI lab as of yet, since at the time of current writing, detecting subtle changes in brain activity as, e.g., speakers name different shades of blue, is likely to be too fine-grained. Despite current limitations, using model-based fMRI decoding to construct a similarity space of semantic representations provides the first steps towards this goal. A straightforward hypothesis would be that one could establish a correlation between the semantic space of, say, color space and the semantic space of color concepts derived from model-based fMRI decoding. As research builds on the results of Zinszer et al. (2016) on ways semantic representations of nominal concepts can remain constant or vary across languages, we may make further progress towards this goal.

Overall, it is clear that there are several tractable questions within linguistics and cognitive neuroscience that can best be answered through interdisciplinary collaboration between the fields. Kemmerer's (2016) topic of nominal classification systems is one such area. Semantic composition (e.g. Shay and Raizada (2015) and Fyshe et al. (2014)), neural decoding Anderson et al. (e.g. 2016a); Zinszer et al. (e.g. 2016), and semantic typology are other feasible areas of research. Importantly, all of these questions would benefit from an increase in the level of interdisciplinary contact between linguistics and cognitive neuroscience. Each field can contribute unique knowledge to the complete understanding of language. Linguistics can provide a deep understanding of complete languages, cross-linguistic similarities and differences, and computational-level models of how language may work. Cognitive neuroscience can provide a deep understanding of how to probe questions about what the human brain can actually do in language, getting at the implementational level. Neither field can answer all questions related to language alone, but combined the fields can develop a more complete, interdisciplinary understanding.

References

Aikhenvald, A. Y. (2000). *Classifiers: A Typology of Noun Categorization Devices: A Typology of Noun Categorization Devices*. OUP Oxford.

Anderson, A. J., Binder, J. R., Fernandino, L., Humphries, C. J., Conant, L. L., Aguilar, M., Wang, X., Doko, D., and Raizada, R. D. S. (2016a). Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation. *Cereb Cortex*. http://dx.doi.org/10.1093/cercor/bhw240

Anderson, A. J., Zinszer, B. D., and Raizada, R. D. S. (2016b). Representational similarity encoding for fMRI: Pattern-based synthesis to predict brain activity using stimulus-model-similarities. *NeuroImage*, 128:44–53. http://dx.doi.org/10.1016/j.neuroimage.2015.12.035

Baroni, M., Bernardi, R., & Zamparelli R. (2014). Frege in space: A program for compositional distributional semantics. In C. Condoravdi, V. de Paiva, & A Zaenen (Eds.), *Linguistic Issues in Language Technology, 9*, 5-110.

Berlin, B. and Kay, P. (1991). *Basic color terms: Their universality and evolution*. Univ of California Press.

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., and Desai, R. H. (2016). Toward a brain-based componential semantic representation. *Cogn Neuropsychol*. Advance Online Publication. http://dx.doi.org/10.1080/02643294.2016.1147426.

Bohnemeyer, J. and Stolz, C. (2006). Spatial reference in Yukatek Maya: A survey. In Levinson, S. C. and Wilkins, D. P., editors, *Grammars of space*, pages 273–310.

Bowerman, M. (1996). The origins of children's spatial semantic categories: Cognitive versus linguistic determinants. In Gumperz, J. J. and Levinson, S. C., editors, *Rethinking linguistic relativity*, pages 145–176. Cambridge University Press, Cambridge, UK.

Contini-Morava, E. and Kilarski, M. (2013). Functions of nominal classification. *Language Sciences*, 40:263–299. http://dx.doi.org/10.1016/j.langsci.2013.03.002

Dowty, D. R., Wall, R. E., and Peters, S. (1981). *Introduction to Montague semantics*, volume v. 11. D. Reidel Pub. Co., Dordrecht, Holland.

Fernandino, L., Binder, J. R., Desai, R. H., Pendl, S. L., Humphries, C. J., Gross, W. L., Conant, L. L., and Seidenberg, M. S. (2015). Concept representation reflects multimodal abstraction: A framework for embodied semantics. *Cereb Cortex*. http://dx.doi.org/10.1093/cercor/bhv020

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. In *Studies in Linguistic Analysis*. Blackwell, Oxford.

Frankland, S. M. and Greene, J. D. (2015). An architecture for encoding sentence meaning in left mid-superior temporal cortex. *Proceedings of the National Academy of Sciences*, 112 VN -(37):11732–11737. http://dx.doi.org/10.1073/pnas.142136112

Fyshe, A., Talukdar, P. P., Murphy, B., and Mitchell, T. M. (2014). Interpretable semantic vectors from a joint model of brain- and text-based meaning. *Proc Conf Assoc Comput Linguist Meet*, 2014:489–499.

Kemmerer, D. (2016). Categories of object concepts across languages and brains: The relevance of nominal classification systems to cognitive neuroscience. *Language, Cognition & Neuroscience*. http://dx.doi.org/10.1080/23273798.2016.1198819

Levinson, S., Meira, S., Language, T., and Group, C. (2003). 'Natural concepts' in the spatial topological domain-adpositional meanings in crosslinguistic perspective: An exercise in semantic typology. *Language*, pages 485–516.

Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information*. W.H. Freeman, San Francisco.

McRae, K., Cree, G. S., Seidenberg, M. S., and McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4):547–559. http://dx.doi.org/10.3758/BF03192726

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., and Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195. http://dx.doi.org/10.1126/science.1152876

Poeppel, D. and Embick, D. (2005). Defining the relation between linguistics and neuroscience. In Cutler, A., editor, *Twenty-First Century Psycholinguistics: Four Cornerstones*, chapter 6, pages 103–118. Psychology Press.

Pulvermüller, F. and Fadiga, L. (2010). Active perception: sensorimotor circuits as a cortical basis for language. *Nature Reviews Neuroscience*, *11*(5), 351–60. http://dx.doi.org/10.1038/nrn2811

Shay, E. A. and Raizada, R. D. S. (2015). Using neurobiologically-motivated features to investigate the semantic composition of adjectives with nouns. Society for Computers in Psychology.

Speed, L., Chen, J., Huettig, F., and Majid, A. (2016). Do classifier categories affect or reflect object concepts? *38th Annual Meeting of the Cognitive Science Society*, page 393.

Turney, P. D. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.

Zinszer, B. D., Anderson, A. J., Kang, O., Wheatley, T., and Raizada, R. D. S. (2016). Semantic structural alignment of neural representational spaces enables translation between English and Chinese words. *J Cogn Neurosci*. Advance Online Publication. http://dx.doi.org/10.1162/jocn_a_01000