How Language Affects Thought in a Connectionist Model

Katia Dilkina (knd@{andrew.cmu.edu, stanford.edu})
Stanford University, Department of Psychology
Jordan Hall (Bldg 420), 450 Serra Mall, Stanford, CA 94305 USA

James L. McClelland (jlm@psych.stanford.edu)
Stanford University, Department of Psychology
Jordan Hall (Bldg 420), 450 Serra Mall, Stanford, CA 94305 USA

Lera Boroditsky (lera@psych.stanford.edu)
Stanford University, Department of Psychology
Jordan Hall (Bldg 420), 450 Serra Mall, Stanford, CA 94305 USA

Abstract

How do the languages we speak shape the way we think? In a series of studies, Boroditsky, Schmidt, and Phillips (2003) investigated the effect of grammatical gender on people's responses to questions about the properties and similarity relations among objects. Here, we use a connectionist network to simulate these findings and find a possible mechanism for linguistic relativity effects (such as effects of grammatical categorization). The model's behavior paralleled the effects seen in the human data. The results also suggest that the within- and between- category similarity relations among objects may play a role in generating these effects.

Keywords: semantics; conceptual knowledge; language; language and thought; linguistic relativity; connectionism.

Introduction

How do the languages we speak shape the way we think? The hypothesis that aspects of language may influence the way we think is most strongly associated with Sapir (1921) and Whorf (1956). Recently, evidence supporting this hypothesis has come from experimental studies documenting that speakers of different languages perform differently on (non-linguistic) tasks such as categorization, perceptual discrimination, and similarity judgments (e.g., Boroditsky, 2001; Boroditsky, Schmidt, & Phillips, 2003; Davidoff, Davies & Roberson, 1999, 2000; Levinson, 1996; Lucy, 1992; Sera et al., 2002; Slobin, 1992, 1996).

In this paper we ask how cross-linguistic differences may arise, and attempt to provide a computational mechanism through which aspects of linguistic representations may influence thinking. In our view, cross-linguistic differences arise as a result of differences in experience. Speakers of various languages are exposed to distinct patterns of linguistic input characterized by its specific statistical properties and associations with other kinds of information. Throughout development, conceptual knowledge about objects in the world is acquired by integrating different kinds of sensory-motor (including linguistic) input. Within this framework, linguistic information about an entity (its name, grammatical gender, other relevant grammatical and syntactic markings, etc.) is treated the same way as other applicable kinds of information – information about what that entity looks like (visual), what it sounds like (auditory), what it feels like (tactile), how it moves (motoric), and so on. The semantic system involves a large network of modality-specific distributed representations (as suggested by imaging studies; Martin & Chao, 2001). Importantly, we believe that there is an additional representation sensitive to both within- and between-modality covariation that serves to link the modality-specific information together (Damasio, 1989; Rogers et al., 2004). These representations combine sensory-motor information with linguistic information and provide the substrate where effects of meaning similarity in picture and word naming and recognition tasks as well as linguistic relativity findings arise (for a related perspective see Vigliocco and colleagues, 2003, 2004, 2005, 2006).

We have used a connectionist implementation of this theory to account for the relationship of semantic and lexical deficits in semantic dementia patients (Dilkina & McClelland, 2006). Furthermore, the theory has implications for a wide range of research areas including language comprehension and production, bilingualism, and conceptual representations and processing (note that for the purposes of this paper, *semantics* is synonymous with *conceptual knowledge*; it does not refer to the meaning of words per se). In the current paper, we will focus on its implications for findings of linguistic relativity, and more specifically on effects of grammatical gender categories.

Grammatical Gender

In English, only persons are referred to with the gender-marked pronouns *he* and *she*. However, in many other languages, including Spanish and German, all nouns are marked for gender – even nouns referring to inanimate objects. For example, the word for 'key' is feminine in

Spanish and masculine in German, while the word for 'bridge' is masculine in Spanish and feminine in German.

In a series of experiments with English-Spanish and English-German bilinguals, Boroditsky et al. (2003) looked at grammatical categorization effects using a set of objects half of which were feminine in Spanish but masculine in German, while the other half were masculine in Spanish and feminine in German.

In one of the experiments, they presented subjects with the English names of the objects and asked them to list the first three adjectives that came to mind to describe each object. The adjectives were scored as describing a feminine or masculine property of the object by five independent naïve native English speakers. The results showed that the adjectives people listed were consistent with the grammatical gender of the objects in their native language.

In another experiment, the investigators presented their subjects with pairs of drawings – one of the items was an object from the list while the other was a male or female person (e.g., a boy vs. a girl). The task was to rate the similarity of each pair on a scale from 1 to 9. Each object was paired with each of eight person drawings. People indicated pairs as more similar when the grammatical gender of the pictured object and the biological gender of the pictured person were consistent than when they were not. Similar results were obtained when the experiment was performed with verbal shadowing where the participants repeated English letters played one per second. Shadowing was included to prevent or at least suppress verbalization.

Boroditsky et al. (2003) found that the *grammatical* category of the labels affected the *semantic* representation of the objects, even though the participants had no idea of the relevance of gender in the tasks. That was the case even when the task involved English labels, English being a language without grammatical gender (as in the first experiment), or non-linguistic stimuli (as in the second experiment). Also, that was true even when linguistic processing of the stimuli was suppressed (as in the third experiment), suggesting that the effects arise at a post-lexical level of representation.

Toward a Mechanistic Account

The goal of the current project was to investigate *the mechanism* by which linguistic relativity phenomena such as gender categorization effects arise. We simulated Boroditsky et al.'s data using a connectionist model. Our findings are discussed with respect to the inherent properties of connectionist networks, namely distributed overlapping representations and sensitivity to coherent covariation, and what they can tell us about the underlying mechanism.

Furthermore, we were interested to see what aspects of the items' perceptual representations are important for gender categorization effects to occur. The two aspects we investigated were the density of the object perceptual representations and the between-category similarity of the persons' and the objects' representations. We hypothesized that as the between-category similarity increased, so would the gender categorization effects. The reason is because the objects' semantic representations are influenced by the perceptual features they share with humans. We also expected that the density of the object representations may also affect the magnitude of the gender effects, or it may modulate the effect of between-category similarity.

Our findings supported the notion that *the language(s) we hear and speak make an important contribution to the representations and organization of our semantic system*. And also, *the similarity structure of pre-semantic representations may play a role in linguistic relativity effects, and clearly affects semantic representations*.

Methods

Network Architecture. The network overall architecture can be seen on Figure 1. There are five visible layers – one for perceptual representations (equivalent to drawings), one for descriptive representations (equivalent to adjectives), and three for the names of the items – in English, Spanish, and German. In the present investigation, each network "speaks" English, and only one other language; thus, only two of the lexical layers are used in a given simulation. The visible layers are bidirectionally connected to a single hidden layer called *semantics*. All layers are self-connected.

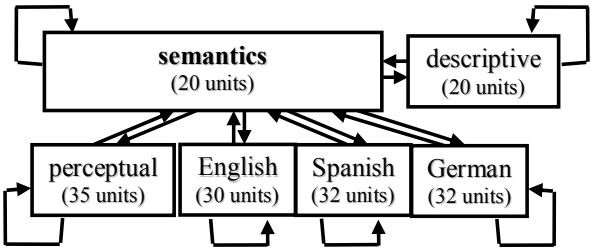

Figure 1: Network architecture.

Simulation Materials. Training and testing patterns were identical. The set included 30 items, 10 persons (5 male, 5 female) and 20 objects (5 feminine in both Spanish and German, abbreviated as **FF**; 5 masculine in both, **MM**; 5 feminine in Spanish but masculine in German, **FM**; and 5 masculine in Spanish, feminine in German, **MF**).

The lexical patterns were localist word representations. Each item had a specific unit as its label in each language. In addition, the Spanish and German representations included two units to mark grammatical gender.

The perceptual and descriptive input patterns were binary patterns generated based on the stochastic prototypes shown in Table 1. There are no necessary or sufficient features included. Rather, these patterns reflect tendencies of entities of particular types (e.g., men and women) to have some properties and not others. Each symbol in the prototypes

stands for the probability of assigning a value of 1 to that vector position. The symbol-probability pairs are included in the table. The positions marked with 'X' and 'Y' varied their probability based on the experimental manipulations, as will be explained shortly. The generated patterns were manually adjusted to ensure equal within- and between-category similarity for each of the four sets of objects.

The perceptual vectors had 35 positions. The first 10 were used to represent predominantly human characteristics; the next 15 object characteristics, and the last 10 correlated with biological gender (5 female, 5 male). The descriptive vectors had 20 positions. The first 10 were more likely to apply to females, the second 10 to males.

Table 1: Perceptual and descriptive prototypes.

perceptual prototypes			
female	++++++++++ ----------	*****	^^^^^
male	++++++++++ ----------	^^^^^	*****
object(low)	------XXXX ==========YYYY	^^^^^	^^^^^
object(high)	------XXXX ******YYYY-----	^^^^^	^^^^^

descriptive prototypes		**symbol**	**P(1)**
female	********** ^^^^^^^^^^	-	0.0
male	^^^^^^^^^^ **********	^	0.2
object	========== ==========	=	0.4
		*	0.6
		+	0.8

Network Training. As mentioned earlier, each network spoke English and either Spanish or German. Thus, there were four relevant patterns for each item – perceptual, descriptive, English, and Spanish or German. During training, the network was given all four, just the perceptual, just the English, or just the Spanish or German pattern, and was asked to produce all four outputs. All items were included in the training in random order. Back-propagation was used to do online learning, where the connection weights between units were updated after every example. The presentation of each example lasted for seven time intervals. During the first three intervals, the input pattern(s) was clamped on. For the remaining four intervals, the input was removed, and the network was allowed to adjust the activation of all units in all layers, including the one(s) previously clamped. During the final two intervals, the unit activations are compared to their corresponding targets. The network was trained with 250 sweeps through the training set, using standard gradient descent with no momentum, with a learning rate of .005 and weight decay of .000005.

Network Testing. To simulate the first paradigm, we used the English patterns of the FM and the MF objects as input. The three most active units in the descriptive layer were selected as indicating the three adjectives the network listed for that item. As in the original Boroditsky et al. experiment, a masculine adjective was given a score of -1, while a feminine adjective was given a score of +1.

To simulate the second paradigm, the network was presented with the perceptual patterns of the FM and the MF objects as well as the persons (female and male). Each item produced a unique semantic activation. For each object-person pair, we used the cosine of their semantic vectors as a measure of similarity. To make it comparable to the nine-point Likert scale used in the experiment, we multiplied the cosine measure by 10 and rounded the number.

To simulate verbal interference, all lexical layers were removed and the network was tested again as just described.

Experimental Design. We manipulated two parameters independently – the within-category similarity of the object perceptual representations and their between-category similarity with person perceptual representations.

In a set of binary representations of fixed size as we have here, the within-category similarity is closely related to the density of the representations, i.e. the number of 1s in a vector given the length of the vector. We chose two levels of object pattern density – *low density of 40%* (six 1s in the 15 positions representing object perceptual characteristics) vs. *high density of 60%* (six 1s in 10 of those 15 positions; the remaining 5 positions were always 0s). The resulting prototypes can be seen in Table 1. Among the actual patterns, the cosine measure of within-category similarity was *.33* for the low-density set and *.46* for the high-density.

Secondly, to vary the between-category similarity, we manipulated the degree of overlap between the persons' and the objects' perceptual prototypes. Initially, they only overlapped in the last 10 positions (representing biological gender). To increase the overlap, we simply shifted the 15 position representing object characteristics forward (so that they overlapped with the first 10 positions representing human characteristics). We shifted them one, two, three, or four times (as indicated by the *X*s in the prototypes). The gap left after each shift is marked with *Y*s in the prototypes and was filled with '–' (i.e. a zero in the patterns). Importantly, this manipulation was done with the actual object patterns, not the prototypes. Therefore, the patterns used were exactly the same at all five levels of overlap. The two categories overlapped in 0, 1, 2, 3, or 4 positions, which translated to a cosine measure of *.08*, *.13*, *.18*, *.23*, and *.28* between each of the four object sets (FF, MM, FM, and MF) and each of the two person sets (female and male); i.e. this was a linear manipulation of between-category similarity.

In summary, this was a 2 x 5 full factorial design. To ensure appropriate sampling, each type of network – Spanish- or German- speaking, with its particular level of pattern density and overlap – was trained 40 independent times (using different random number generator seeds) and tested for the three experiments. Once again, the only difference between the Spanish- and the German- speaking networks was the grammatical gender of the FM and MF objects. All other patterns were identical.

Results

Experiment 1. Our results mirrored the findings of Boroditsky et al. Language was not a significant factor ($F(1,78)=1.796$, $p=.184$), but it interacted with the item type ($F(1,78)=314.6$, $p<.0005$). As can be seen on Figure 2, the *genderedness* of the object descriptors the networks provided was consistent with the *grammatical gender* of the objects in their "native" language. For the FM subset of objects, which were feminine in Spanish and masculine in German, the Spanish-speaking networks described these objects as more feminine than the German-speaking networks; and vice versa for the MF objects. Furthermore, the density of the object perceptual representations modulated this interaction ($F(1,78)=17.80$, $p<.0005$), while the object-person pattern overlap did not ($F(4,312)=1.097$, $p=.358$).

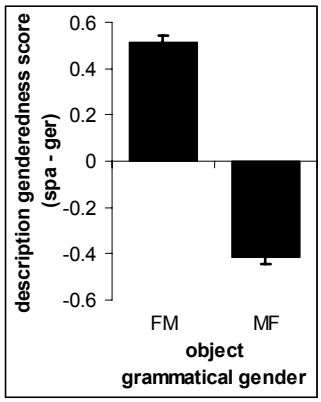

Figure 2: Language by gender interaction in Exp 1.

FM objects = feminine in Spanish, masculine in German; MF objects = masculine in Spanish, feminine in German; The y-axis shows the difference in the description genderedness between Spanish- and German- speaking networks (positive = more feminine, negative = more masculine).

Experiment 2. Again, language was not a significant factor ($F(1,78)=1.209$, $p=.314$), but just as in Boroditsky et al.'s study there was a three-way interaction between L1, object gender, and person gender ($F(1,78)=1969.9$, $p<.0005$). The networks exhibited higher semantic similarity for pairs where the *grammatical gender* of the object and the *biological gender* of the person were consistent than for pairs where they were not. For example, as can be seen on Figure 3a,when FM objects (feminine in Spanish, masculine in German) were paired with female persons, Spanish-speaking nets indicated a higher semantic similarity than German-speaking ones. Conversely, when the same objects were paired with male persons, the German-speaking nets indicated higher semantic similarity than Spanish-speaking ones. The object-person pattern overlap significantly modulated this interaction ($F(4,312)=9.209$, $p<0.0005$); the density of the object representations did so marginally ($F(1,78)=3.179$, $p=.078$); and there was no five-way interaction ($F(4,312)=1.42$, $p=.228$). An increase in overlap or in density generally resulted in an increase in the effect observed (Figure 4). Finally, there was no main effect of density ($F(1,78)=.649$, $p=.423$), but there was a main effect of overlap ($F(4,312)=9.91$, $p<.0005$) such that as the degree of overlap increased so did the average similarity rating.

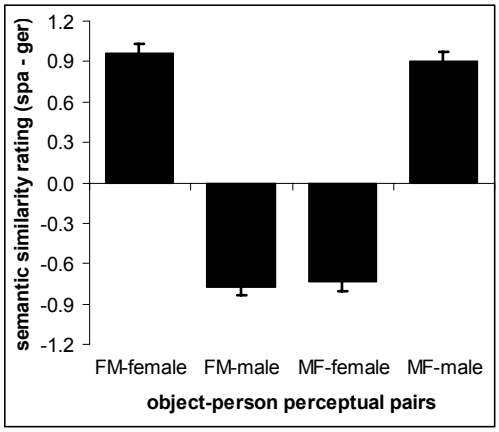

Figure 3a: L1 by object gender by person gender in Exp 2.

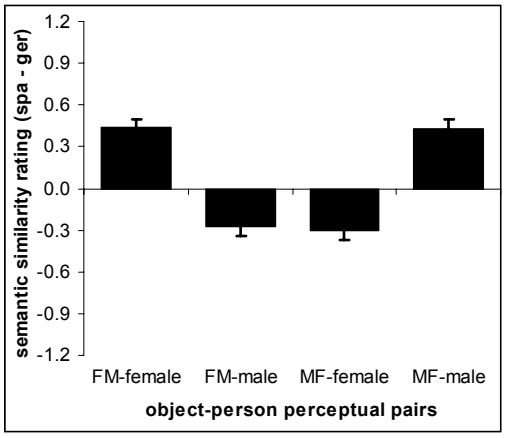

Figure 3b: L1 by object gender by person gender in Exp 3.

FM objects = feminine in Spanish, masculine in German; MF objects = masculine in Spanish, feminine in German; The y-axis shows the difference in semantic similarity between Spanish- and German- speaking networks.

Experiment 3. The results were similar to Experiment 2, although the effects were weaker. The three-way interaction between L1, object gender, and person gender ($F(1,78)=1276.6$, $p<.0005$) is presented on Figure 3b. The modulation of this interaction by object-person pattern overlap was not significant ($F(4,312)=.818$, $p=.514$); and there was a significant modulation of the interaction by density ($F(1,78)=4.044$, $p=.048$). Both can be seen on Figure 4. Again, there was no main effect of L1 ($F(1,78)=.68$, $p=.412$) or density ($F(1,78)=.056$, $p=.813$), but a main effect of overlap ($F(4,312)=24.46$, $p<.0005$).

Discussion

The results we obtained were consistent with the findings of Boroditsky et al.: (1) In the object description paradigm, there was a significant language by item interaction indicating that the networks' description of a given item was masculine- or feminine- biased depending on the language the network "spoke", and thus on the grammatical gender of the item in that language. (2) In the object-person picture similarity paradigm, there was a significant language by object gender by person gender interaction indicating that the networks' similarity ratings were higher when the object's grammatical gender and the person's biological gender were consistent (both feminine, or both masculine) than when they were not (one feminine, one masculine). (3) This interaction persisted and was still highly significant even under an extreme manipulation of verbal interference, that is when all lexical layers of the networks were removed. This finding supported our view that *linguistic information helps shape semantic representations throughout development* (or throughout training for the network), and the driving cause of the observed effects is not merely the participants' verbalizing in their native language or the networks' activating its lexical representations.

It should be noted that the effect decreased in size under verbal interference (Figure 3). Thus, though the grammatical categorization effects arose at the semantic level, there is also on-line contribution from the lexical representations – which when intact boost up the saliency of grammatical gender. However, this contribution is not critical.

In the original studies by Boroditsky et al., there was no difference in the effect size between the verbal-interference experiment and the no-interference one. One possibility is that verbalization may not be necessary for lexical access, and so the verbal suppression did not interfere with the online lexical support in the behavioral experiments. With respect to the simulation results presented here, the observed difference between Experiment 2 and 3 is most likely due to the extreme implementation of verbal interference. A more moderate implementation would have been to introduce noise at the lexical levels rather than shut them down entirely. However, we felt it is important to show that the effects can be found even in this most extreme situation so that we can confidently draw the conclusion that the effects arise at the semantic level.

The more interesting question we set out to answer was about *the mechanism* underlying the observed effects. There are two attributes of connectionist networks that are relevant here, and importantly, neither of them is tied to the particular architecture or size of the network. They are general properties of this type of networks. The first one is that they are sensitive to the *coherent covariation* of the various properties of items. We emphasize here that linguistic properties should be included. For example, the biological gender of a person and the grammatical gender used when referring to that person are perfectly correlated. In addition, the biological gender of a person and the femininity or masculinity of the adjectives we use to describe them are also correlated (even if not perfectly). The network learns the coherent covariation between perceptual, descriptive, and grammatical properties of humans.

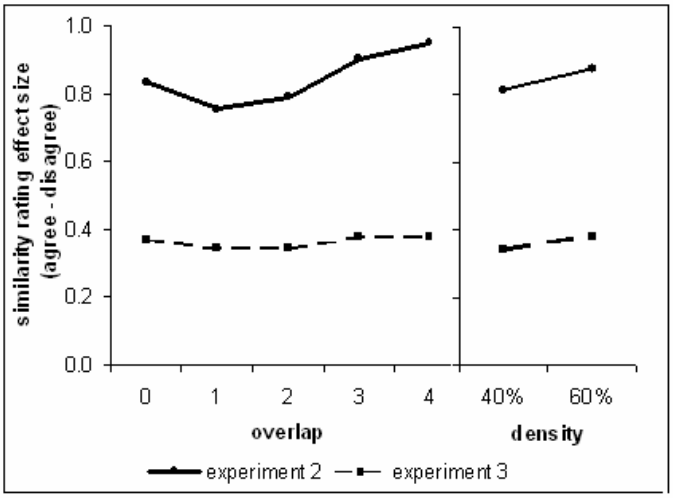

Figure 4: Effects of overlap and density on the L1 by object gender by person gender interaction in Exp 2 & 3.

The y-axis plots the difference between semantic similarity of object-person pairs where the object and the person agree on grammatical gender vs. pairs where they do not.

What about the objects? They have no biological gender and their descriptors are generally not biased. Why does their grammatical gender affect their semantic representations if there is no coherent covariation between perceptual, descriptive, and lexical properties? The second important attribute of connectionist networks is that they use *distributed overlapping representations*. Because of that, all categories "borrow" partial sensitivity to coherent covariation, even when it is completely unsupported in a category. The objects, therefore, borrow from the structure of the person representations and exhibit the respective bias.

In relation to this second property, we postulated that increasing the overlap among objects' and persons' perceptual patterns will result in a stronger effect. Why? As the between-category similarity increases, the semantic representations of the objects would be more similar to the semantic representations of the persons, which in turn are based on the coherent covariation among the persons input patterns (perceptual, descriptive, and lexical). That coherent covariation demands that there is a relationship between the grammatical gender of an item and other characteristics such as what the item looks like or how it is described. Indeed, the results from the second experiment confirmed that (a) the *semantic similarity* between objects and persons increased with overlap in the perceptual representations; and

(b) that increase was larger for object-person pairs that agreed on gender than pairs that did not (Figure 4).

Importantly, the overlap we explored involved not features of biological gender but overall person vs. non-person distinctions. The motivation was that there are different kinds of non-person categories. For example, animals share many more features with humans than artifacts do; within artifacts, toys may share more features with humans than household items do; etc. Our findings suggest that if the object-person picture similarity paradigm was tested using a set of animals instead of a set of objects, the grammatical gender effect should be even stronger. This is partially supported by Vigliocco et al.'s study (2005), where animals that shared grammatical gender were found to be more similar to each other than those that did not; while no such difference was present for artifacts.

As mentioned, the positions of overlap had nothing to do with the male vs. female distinction. Thus, the increase in overlap promoted the object category to "borrow" more partial sensitivity to coherent covariation from the person category, but it did not lend direct support for it. Because of that, the interaction seen in Experiment 2 disappeared in Experiment 3, where the lexical layers (which do in fact lend direct support) were not allowed to contribute to the semantic activation. This also emphasizes the idea that two mechanisms seem to be at work: (1) *a long-term mechanism whereby linguistic information helps shape conceptual representations*; and (2) *an on-line mechanism whereby the activation of lexical representations bring linguistic information to bear on conceptual tasks*.

A natural extension of this project is to test the effect of increased overlap in the perceptual features marking *biological gender*. Our prediction would be that the strengthening of the grammatical gender effects would be even more pronounced than increasing the overlap among the other perceptual features.

References

Boroditsky, L. (2001). Does language shape thought? English and Mandarin speakers' conceptions of time. *Cognitive Psychology, 43*, 1-22.

Boroditsky, L., Schmidt, L., & Phillips, W. (2003). Sex, Syntax, and Semantics. In Gentner & Goldin-Meadow (Eds.), *Language in Mind: Advances in the Study of Language and Cognition*. MIT Press: Cambridge, MA.

Damasio, A.R. (1989). The brain binds entities and events by multiregional activation from convergence zones. *Neural Computation, 1*, 123-132.

Davidoff, J., Davies, I., & Roberson, D. (1999). Colour categories in a stone-age tribe. *Nature, 398*, 203-204.

Dilkina, K., & McClelland, J.L. (2006). A connectionist account of the pattern of deficits across semantic and lexical tasks in five semantic dementia patients. Poster presented at the 28th Annual Meeting of the Cognitive Science Society, Vancouver, BC.

Levinson, S.C. (1996). Frames of reference and Molyneux's question: Crosslinguistic evidence. In P. Bloom, M. Peterson, L. Nadel, & M. Garrett (Eds.), *Language and space*. Cambridge, MA: MIT Press.

Lucy, J.A. (1992). *Grammatical Categories and Cognition: A Case Study of the Linguistic Relativity Hypothesis*. Cambridge, UK: Cambridge University Press.

Martin, A., & Chao, L.L. (2001). Semantic memory and the brain: Structure and processes. *Current Opinion in Neurobiology, 11*, 194-201.

Roberson, D., Davies I., & Davidoff, J. (2000). Color categories are not universal: Replications and new evidence from a Stone-age culture. *Journal of Experimental Psychology: General, 129*, 369-398.

Rogers, T.T., Lambon-Ralph, M.A., Garrard, P., Bozeat, S., McClelland, J.L., Hodges, J.R., & Patterson, K. (2004). The structure and deterioration of semantic memory: A neuropsychological and computational investigation. *Psychological Review*, *111*, 205-235.

Sapir, E. (1921). *Language*. New York, NY: Harcourt, Brace, and World.

Sera, M., Elieff, C., Forbes, J., Burch, M.C., Rodriguez, W., Dubois, D.P. (2002). When language affects cognition and when it does not: An analysis of grammatical gender and classification. *Journal of Experimental Psychology: General, 131*, 377-397.

Slobin, D.I. (1992). *The Cross-Linguistic Study of Language Acquisition*. Hillsdale, NJ: Erlbaum.

Slobin, D.I. (1996). From "thought and language" to "thinking for speaking". In J. Gumperz & S. Levinson (Eds.), *Rethinking Linguistic Relativity*. Cambridge, MA: Cambridge University Press.

Vigliocco, G., & Kita, S. (2006). Language-specific properties of the lexicon: Implications for learning and processing. *Language & Cognitive Processes,21*,790-816.

Vigliocco, G., Vinson, D.P., Lewis, W., & Garrett, M.F. (2004). Representing the meaning of object and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology, 48*, 422-488.

Vigliocco, G., Vinson, D.P., Paganelli, F., & Dworzynski, K. (2005). Grammatical gender effects on cognition: Implications for language learning and language use. *Journal of Experimental Psychology, 134*, 501-520.

Vinson, D.P., Vigliocco, G., Cappa, S.F., & Siri, S. (2003). The breakdown of semantic knowledge: Insights from a statistical model of meaning representation. *Brain & Language, 86*, 347-442.

Whorf, B. (1956). In J. B. Carroll (ed.), *Language, Thought, and Reality: Selected Writing of Benjamin Lee Whorf*. Cambridge, MA: MIT Press.