An Empirical and Computational Test of Linguistic Relativity

Kathleen M. Eberhard* (eberhard.1@nd.edu)
Matthias Scheutz (mscheutz@cse.nd.edu)**
Michael Heilman (mheilman@nd.edu)**
*Department of Psychology, **Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN 46556 USA

Abstract

To what extent does the correlation between grammatical gender and conceptual sex in many languages result in speakers having an implicit association between sex and the concepts of inanimate objects? This question was examined in an artificial gender-learning task similar to Phillips and Boroditsky (2003). The task required native English speakers to learn the grammatical gender of nouns denoting inanimate objects (e.g., a fork) as well as humans (e.g., a man). The speakers then rated the similarity of pictures of the inanimate objects and the humans. Consistent with Phillips and Boroditsky's results, speakers rated an object and human as more similar when their nouns' gender was consistent than when it was inconsistent. Furthermore, this consistency effect occurred for objects that were paired with pictures of humans in which no explicit association of gender had been learned. A connectionist model tested hypotheses about the associative links that underlie the consistency effects in the ratings as well as how the speed of learning affects those associations. Together the empirical data and the model simulations demonstrate that associative connections between inanimate object concepts and conceptual properties of sex are unnecessary for the consistency effects.

Introduction

According to linguistic relativity, differences in the vocabulary and grammar of languages cause speakers to conceptualize the world differently. Previous empirical tests of this view often focused on whether speakers of two different languages differ in their ability to distinguish objects within a category (e.g., colors) when the languages differ in the size of the vocabulary used to refer to that category. However, several recent studies have examined the influence of language on thought by investigating whether speakers of languages with grammatical gender have implicit associations between concepts of inanimate objects and the conceptual properties of male and female sex as a result of a correlation between the grammatical gender and sex (e.g., Boroditsky, Schmidt, & Phillips, 2003; Phillips & Boroditsky, 2003; Sera, Elieff, Forbes, Burch, Rodriguez, & Dubois, 2002; Vigliocco, Vinson, & Paganelli, 2004).

In particular, Boroditsky, Schmidt, & Phillips (2003; Phillips & Boroditsky, 2003) conducted a set of studies showing that speakers of languages such as German and Spanish conceptualize inanimate objects denoted by masculine nouns as being male-like and inanimate objects denoted by feminine nouns as being female-like. For example, in one experiment, Boroditsky et al. presented pairs of pictures consisting of a female or male, such as a bride or king, and an inanimate object, such as a spoon or clock, to a group of native German speakers and to a group of native Spanish speakers. Both groups were instructed to rate the similarity of the objects in each pair on a 9-point scale. In both German and Spanish, the gender of a noun denoting a person almost always matches the person's biological sex (e.g., Die Braut$_{\text{feminine}}$/la novia$_{\text{feminine}}$ [*the bride*]; Der König$_{\text{masculine}}$/el rey$_{\text{masculine}}$ [*the king*]). However, the gender of the nouns denoting the inanimate objects in Boroditsky et al.'s experiment was opposite in the two languages. So, for example, the noun denoting a spoon is masculine in German (Der Löffel) but is feminine in Spanish (la cuchara), and the noun denoting a clock is feminine in German (Die Uhr) but is masculine in Spanish (el reloj). The results of Boroditsky et al.'s rating task showed that the two groups of speakers rated inanimate objects as more similar to the human entities when the gender of the two nouns was the same than when the gender was different. Thus, German speakers rated a spoon and a king as more similar than a spoon and a bride whereas the Spanish speakers rated a spoon and bride as more similar than a spoon and a king.

Boroditsky et al. provided additional evidence using an artificial gender-learning task with English speakers, the results of which were further investigated in the study reported here. Specifically, native English speakers were taught an "artificial" language in which nouns were classified as either "soupative" or "oosative". The speakers were told that the classification was reflected in whether a noun is preceded by the article "sou" or "oos". To learn the classification, the speakers were shown 20 pictures of objects along with a label consisting of "sou" or "oos" and the English noun used to refer to the object. Ten pictures were "sou" objects and ten were "oos" objects. The six inanimate objects in one gender set consisted of *pear*, *fork*, *violin*, *pot*, *pen*, and *cup* whereas the six objects in the other gender set consisted of categorically related objects such as *apple*, *spoon*, *guitar*, *pan*, *pencil*, and *bowl*. The four other items in each set of ten were either females (*ballerina*, *bride*, *woman*, and *girl*) or males (*man*, *boy*, *giant*, *king*). Thus, similar to the partial correlation between sex and gender in many natural languages, the artificial language had a partial correlation between sex and grammatical gender in the items, as a result of all of the females being associated with one gender (e.g., "soupative") and all of the males being associated with the other (e.g., "oosative"). The assignment of gender and the association of the males and females with the two sets of inanimate objects were counterbalanced across lists.

During a learning phase in the experiment, the 20 pictures were presented three times, in random order, along with the determiner "oos" or "sou" and the noun referring to the depicted object. After the learning phase, the 20 pictures were presented without their labels, and the speakers' indicated the corresponding gender by pressing a key labeled "oos" or "sou" on a computer keyboard. After the speakers correctly classified the gender of all 20 pictures, they were given a rating task similar to the task given to the German and Spanish speakers. In particular, all eight human pictures were paired with each of the 12 inanimate pictures for a total of 96 pairs. The pairs were again presented without any "oos" or "sou" labels and the speakers rated each pair with respect to the similarity of the human and inanimate object on a 9-point scale. Similar to the rating results from the German and Spanish speakers, the English speakers' rating exhibited a "gender consistency effect", such that higher similarity ratings were given to pairs in which the inanimate object's gender was consistent with the human's gender/sex relative to pairs in which the inanimate object's gender was inconsistent with the human's gender/sex.

The current study was designed to further test the nature of the associations that were responsible for the English speakers similarity ratings. Specifically, if the correlation between sex and gender in the artificial language caused the speakers to form an association between a sex property (e.g., male) and the concept of an inanimate object, such as "fork" that is associated with the correlated gender (e.g., "oosative"), then that generalized conceptual association should lead speakers to rate the picture of an inanimate object, such as fork, as being more similar to a picture of a "new" male human, such as "groom", which does not have an explicitly learned association with gender because it was not presented during the gender-learning phase of the experiment. To verify whether direct connections between the conceptual properties of male and female sex and the concepts of inanimate objects are necessary for consistency effects observed in the ratings of pairs with either "old" humans or "new" humans, connectionist models were constructed to simulate the empirical data.

Experiment

Participants

Twenty-four native English speakers from the University of Notre Dame participated in the study in exchange for extra course credit.

Materials

The materials consisted of 24 pictures; half depicted different categories of humans and half depicted different categories of inanimate objects, with the latter being the same objects that were used in Phillips and Boroditsky's (2003) Experiment 4. The twelve inanimate objects were divided into two sets. One set consisted of *apple*, *spoon*, *guitar*, *pan*, *pencil*, and *bowl*, and the other set consisted of items from the same categories as those in the first set, namely, *pear*, *fork*, *violin*, *pot*, *pen*, and *cup*. The twelve human pictures were also divided into two sets, each consisting of three males and three females. One set consisted of *priest*, *boy*, *king*, *bride*, *woman*, and *grandmother*, and the other set consisted of categorically related humans of the opposite sex: *nun*, *girl*, *queen*, *groom*, *man*, and *grandfather*.

Four lists of 18 pictures were constructed for the gender-learning phases of the experiment. The 18 pictures included both sets of inanimate objects and one set of human pictures. Across all four lists, both sets of human pictures were included in two lists. In each list, half of the items were assigned "oosative gender" and the other half were assigned "soupative gender", with each half consisting of one set of six inanimate objects plus either the three male pictures or the three female pictures. The crossing of the male and female pictures with the two sets of inanimate objects and the assignment of gender were counterbalanced across the four lists.

Two lists of 72 pairs of pictures were constructed for the rating task by pairing each of the human pictures in the two sets with the 12 inanimate object pictures. The order of the 72 pairs in each list was random. Both lists (all 144 pairs) were presented in the rating task, with the list containing pairs with the "old" humans that were presented in the learning phase occurring before the list containing pairs with the "new" humans.

Procedure

Participants were run individually. They were seated in front of a computer in a small quiet room and were told that the experiment investigates people's ability to learn to classify words of an artificial language. The experiment was presented in three phases: learning, test, and rating. During the learning phase, one of the four lists of 18 pictures was presented, with all four lists being presented to an equal number of participants. Each picture was presented three times in the center of a computer screen along with a label consisting of either "oos" or "sou", depending on the picture's gender assignment in the list, and the name of the depicted object (e.g., "oos groom", "sou spoon"). The pictures were presented in random order and were displayed for a duration of three seconds. The test phase tested the participants' learning of the 18 pictures' gender. Specifically, the pictures were presented in random order without their labels. The participants were instructed to indicate whether each picture was an "oos" item or a "sou" item by pressing the appropriately labeled key on the computer's keyboard. Feedback was given after each response by displaying the message "Correct" or "Incorrect" for two seconds on the screen. The list of 18 pictures continued to be presented until the participants made 18 consecutive correct responses to all of the items or had attempted to do so within a maximum of 100 trials. The rating task began immediately after the test phase. The participants were told that 144 pairs of pictures would be presented with each pair consisting of a human and an inanimate object. They were instructed to rate the similarity of the two items on a 9-point scale, with 1 corresponding to not at all similar and 9 corresponding to very similar. The participants were encouraged to use the entire range of the scale. Each pair was presented with the human picture on the left side of the screen, the inanimate object picture on the right side, and the 9-point rating scale with the labeled endpoints at the top of the screen. The entire experimental session lasted approximately 30 minutes.

Results

Two 2X2X2 ANOVAS were conducted on the average similarity ratings, one with participants as a random factor and the other with items as a random factor, designated as $F1$ and $F2$, respectively. The familiarity of the human pictures ("old" (presented during the learning phase) or "new") and the consistency of the inanimate object's gender with the gender/sex of the human were within-participant and within-item factors, and the human's sex (male vs. female) was a within-participants factor but a between-items factor.

Only the main effect of consistency was significant ($F1(1, 23) = 8.20$, $p < .01$; $F2(1, 10) = 107.14$, $p < .001$), with higher average similarity ratings occurring for pairs in which the inanimate object's gender was consistent with the human's gender/sex than for pairs in which the gender was inconsistent. No other main effects nor interactions were significant.

There was variability among the participants with respect to the number of test trials that were required before they reached the criterion for learning the gender of the 18 items during the test phase. Thus, the results were further examined by dividing the participants into two equal groups of "fast learners" and "slow learners". Eleven of the 12 fast learners completed the test phase in the minimum of 18 trials by correctly identifying the gender of all 18 items on the first pass through the list. The other fast learner completed the test phase in 25 trials due to making one incorrect response. The 12 slow learners completed the test phase after an average of 71 trials, with three failing to meet the criterion of 18 consecutive correct responses within the maximum of 100 trials. The slow learners made an average of 14 incorrect responses, most of which occurred with the inanimate objects.

Figure 1 below shows the average similarity ratings for the conditions in which the rating pairs contained an "old" human or a "new" human and the gender of the inanimate object was "consistent" or "inconsistent" with the human's gender/sex. The same ANOVAs were conducted on the average similarity ratings but with the addition of learner (fast or slow) as a between-participants factor and a within-items factor. The main effect of consistency was significant ($F1(1, 22) = 9.15$, $p < .01$; $F2(1, 10) = 107.14$, $p < .0001$). However, the interaction between consistency and learner was marginally significant in the participant analysis ($F1(1, 22) = 3.68$, $p = .07$), and significant in the item analysis ($F2(1, 10) = 19.22$, $p < .01$).

As Figure 1 shows, the fast learners' ratings exhibited a consistency effect regardless of whether the pairs contained an old human or a new human. However, the slow learners did not show a reliable consistency effect for either the pairs with old humans or new humans. No other main effects or interactions were significant.

Discussion

Consistent with Philips and Boroditsky's (2003; Boroditsky, Schmidt, & Phillips, 2003) findings, the results showed that when there is a correlation between grammatical gender and sex, inanimate objects, which have no biological sex, are rated as more similar to humans when the grammatical gender of the object's and human's nouns is consistent than when it is inconsistent. Furthermore, the current study shows that the consistency effect generalizes to similarity ratings for inanimate objects paired with new humans (i.e., humans referred to by nouns in which no prior association with gender was explicitly learned). However, the current study also showed that these consistency effects depended on the rate at which the explicit association of gender with the set of human and inanimate objects was learned. Specifically, participants who quickly learned the association exhibited the consistency effects in the similarity ratings whereas participants who took three or more times longer to learn the associations did not show any reliable consistency effects. To further explore the effect of learning rate on the consistency effects as well as the nature of the associations that underlie the effects, connectionist models were constructed to simulate the empirical data.

Model Simulations

Model Architecture

We distinguish three levels in the model architecture: (1) an visual input level with orthographic and pictorial representations, (2) a lexical-grammatical level consisting of abstract modality independent lexical (word) representations and the grammatical features associated with them, and (3) a level consisting of concepts and conceptual properties. At the orthographic level, words are recognized as letter patterns (e.g., Rumelhart & McClelland, 1986) and are connected to corresponding abstract lexical (word) representations at the lexical-grammatical level. The lexical representations are connected to grammatical features associated with them (e.g., gender). The orthographic nodes corresponding to the artificial determiners "oos" and "sou" are connected to their respective grammatical category nodes for *oosative* and *soupative* gender. The lexical-grammatical nodes are connected to associated concepts. Each concept node receives activation from its associated lexical-grammatical representation and/or directly from its picture input node. The conceptual properties of male and female sex, in turn, receive activation from associated human concept nodes. Because we are only interested in the process of activating sex properties, the model did not include representations for any other conceptual properties.

Figure 2 illustrates a reduced version of the full implemented model, and shows one exemplar for each relevant test condition (i.e., four words from the test set representing each of four possible combinations of "oos" and "sou" with nouns for a male and a female person and two nouns for inanimate objects). Boxes denote computational nodes, and lines with arrows denote directed connections, all of which were excitatory. The dashed lines indicate connections that are not present in the model before training, but which might form as a result of the learning process. In particular, there were four sets of learnable connections: *lexical*-*gender* connections (e.g., between the *spoon* lexical node and the *oosative* gender node), *gender*-SEX connections (e.g., between the *oosative* gender node and the MALE node), CONCEPT-*gender* connections (e.g., between the SPOON concept node and the *oosative* gender node), and inanimate CONCEPT-SEX connections (e.g., between the SPOON concept node and the MALE node). The aim is to test the simplest model, and, thus, there are only a few parameters, of which all but one are fixed.

The computational units were simplified versions of the interactive activation and competition units used for word recognition (e.g., Rumelhart & McClelland, 1986), whose change in activation is given by

$$\Delta act/\Delta t = netin - act \bullet (netin + decay)$$

where *act* is a unit's activation (within [0,1]), *netin* is the summed weighted input to the unit, and *decay* is a constant equal to 0.05 for all nodes. The solid lines depict connections that exist before training, all of which had a fixed maximum excitatory weight of 0.1. For associative learning, the following weight update rule was used, which is a version of Hebbian learning adapted for our computational units:

$$\Delta act/\Delta t = \eta \bullet act_i \bullet act_j \bullet (0.1 - w_{i,j})$$

where $w_{i,j}$ is the connection weight between units i and j, act_i and act_j are the respective units' activations, and η is the learning rate, which is the model's only remaining free variable. Finally, similarity ratings needed to be derived from the model in order to compare it with the participants' ratings in the Experiment. We assumed that the similarity between two given items depends on the number of shared properties that are activated by the items' representations. A picture of a priest and a spoon, for example, will both activate the *oosative* node, but not the *soupative* node; hence "priest" and "spoon" agree with respect to *oosative*. "Priest" also activates "male", but since "spoon" is inanimate it does not activate either "male" or "female", and, therefore, there is no disagreement between the sex property nodes, but no agreement either. Note that "agreement" manifests itself in higher activations of a node as it will receive excitatory input from two processing routes.

The relevant categories are "oosative-soupative" and "male-female". To derive similarity ratings for two items that are presented as input to the model, we define a mapping $F(m,f,o,s) = |m-f| - |o-s| + c$ from node activations (m,f,o,s) to the ratings, which computes the sum of the absolute differences between two conflicting property nodes (c is a constant to scale the quantity to the human ratings).

Simulation Methodology

The main question addressed by the model is whether the gender-learning task causes the formation of the sets of connections depicted by dashed lines in Figure 2. We formulate three hypotheses: (H1) the difference in slow versus fast learners' ratings is due to the difference between the two groups' learning rate; (H2) the *lexical*-*gender* connections as well as the *gender*-SEX connections will form as a result of the gender-learning task and will account for the consistency effects in the ratings; (H3) contrary to Linguistic Relativity explanation of the consistency effects, it is unlikely that additional CONCEPT-*gender* connections and inanimate CONCEPT-SEX connections, which might form during learning, will contribute substantively to the effects.

To test the hypotheses, we first fit the learning rate parameter η to the empirical data such that the model predicts the participants' ratings. If the results are predicted correctly for both the slow and the fast learners, we can then examine the model and trace the flow of activations to determine which connections contributed to the ratings.

The first two hypotheses were tested by constructing two different models, one for slow learners (η = 0.04) and one for fast learners (η = 0.12). We then presented the same training set from the Experiment to both models. As in the learning phase of the Experiment, this set was presented three times in random order with the weights on the *lexical*-*gender* and *gender*-SEX connections being updated after 100 cycles following the onset of an input item. After training, both models were tested on a test set, which was the same as the training set, and were allowed to learn based on feedback about the accuracy of the gender classification of each input item. A threshold-based criterion for correct classification was used, i.e., the activation of the target node (e.g., *oosative*) had to be greater than the classification threshold CT = 0.45, and the activation of the non-target node (e.g., *soupative*) had to be less than an error threshold ET = 0.05.[1] The complete set of 18 test items was presented once to the fast-rate model and four times to the slow-rate model, corresponding to the average number of test trials for the fast and slow learners in the Experiment. Then, both models were run on the rating task, which required deriving a similarity rating for four pairs: "priest-spoon" corresponding to the Old-Consistent gender condition, "priest-fork" corresponding to the Old-Inconsistent gender condition, groom-spoon corresponding to the New-Consistent gender condition, and "groom-fork" corresponding to the New-Inconsistent gender condition. One pair served as input at a time, and when the network settled, the rating was computed based on F as described above.

---

[1] The particular values are not important, only that $CT > ET + c$ for some $c \in [0,1]$ and that the values be fixed in advance (based on the other parameters) and applied to all models.

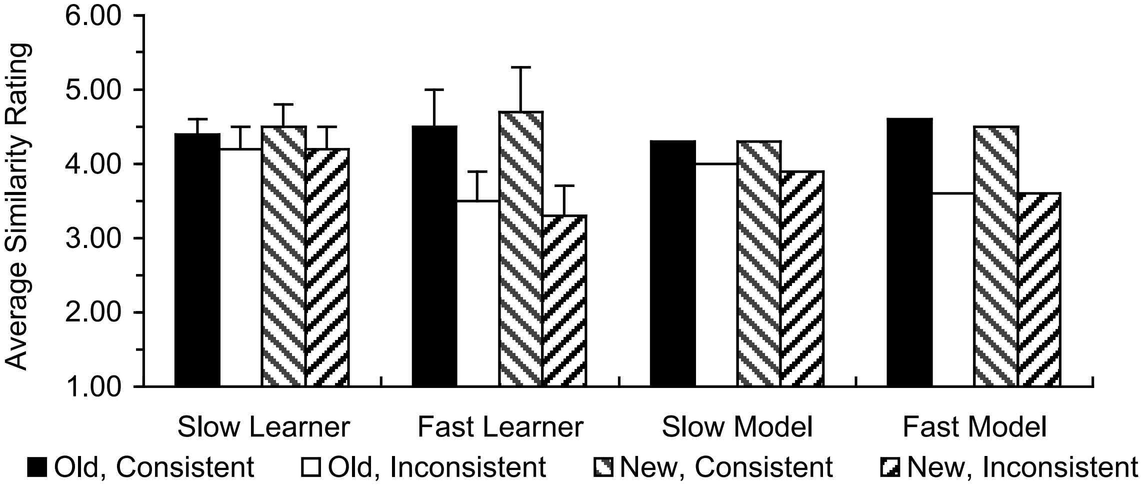

Figure 1: The slow and fast learners' average similarity ratings (and standard errors) for four conditions of picture pairs with objects and humans, and the ratings in the corresponding conditions produced by models trained with a slow or fast learning rate (1 = no similarity; 9 = very similar). Old vs. New refers to whether the human was presented during the gender-learning phase, and Consistent vs. Inconsistent refers to whether the object's gender was consistent with the human's gender/sex.

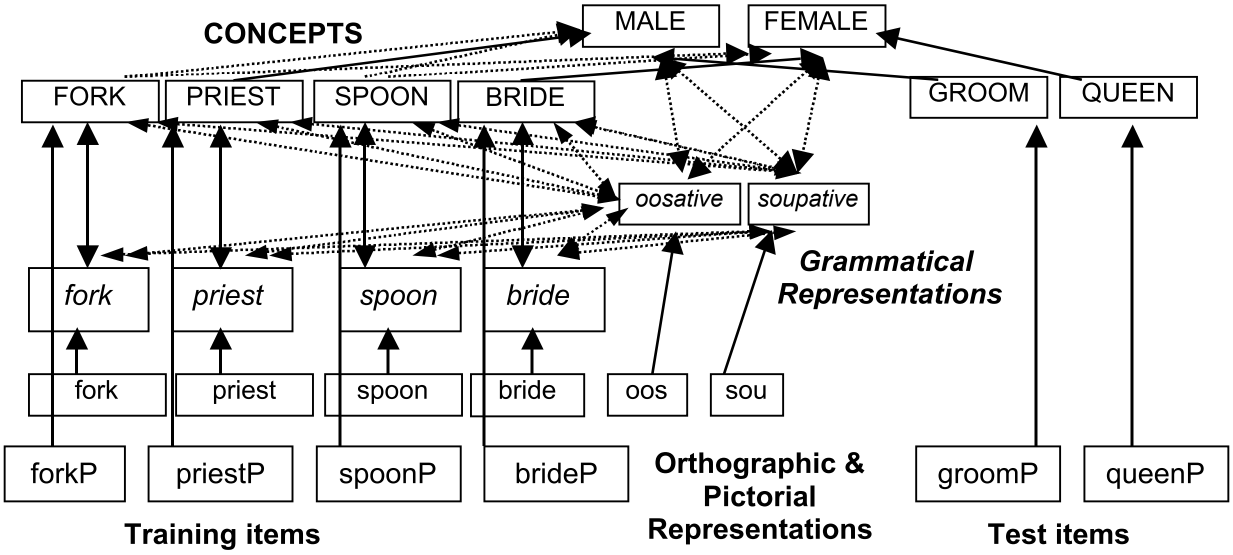

Figure 2: The basic model architecture with pictorial (indicated by P) and orthographic input representations, lexical-grammatical representations (in italics), and concept representations (in caps). Dashed lines depict connections that do not exist before learning but can form as a result of learning.

Simulation Results

As shown in Figure 1, both the slow- and fast-rate models predict the ratings of the slow and fast learners in the Experiment. The overall correlation between the model's ratings and the participants' ratings is 0.92 (the correlation is 0.86 for the slow model and slow learners, and the correlation is 0.96 for the fast model and fast learners). The difference in learning rate accounts for the difference in the two groups' ratings supporting hypothesis (H1). Specifically, both the *lexical-gender* and *gender*-SEX connections formed in the models but with different strengths. If these connections are removed, the models are unable to fit the participants' rating data. In particular, if the *lexical-gender* connections are removed, the models cannot correctly categorize the gender of the training input items. If the *gender*-SEX connections are removed, the fast-rate model no longer produces a consistency effect in the rating for pairs with new humans. Hence, the necessity of these two sets of connections supports hypothesis (H2).

To test the third hypothesis (H3), we repeated the simulations allowing all four sets of connections to form during learning. Both the CONCEPT-*gender* connections and inanimate CONCEPT-SEX connections did form due to the coactivation of the SEX nodes, *gender* nodes, and CONCEPT nodes. The results of the rating tests with these "full models" are essentially the same as the previous models' results: The overall correlation between the full models' ratings and the participants' ratings is 0.89 (the correlation is 0.88 for the slow model and slow learners and is 0.97 for the fast model and fast learners). The critical question was whether the inanimate CONCEPT-SEX connections and/or the CONCEPT-*gender* connections substantively contribute to the consistency effects in the similarity ratings. This question was addressed by examining the correlations without these two sets of connections. Specifically, there was a negligible change in the correlations: The overall correlation between the models' ratings and the participants' ratings is 0.91 (0.87 for the slow model and slow learners, and 0.93 for the fast model and fast learners). Thus, this result confirms hypothesis (H3), that any connections between inanimate concepts and sex properties that form during the learning task are irrelevant to the consistency effects observed in the ratings.

Discussion

The results from the model simulations strongly suggest that the *lexical*-*gender* and *gender*-SEX connections are responsible for the consistency effects in the participants' ratings for several reasons: (1) the *lexical*-*gender* connections form the basis of grammatical categorizations (this is particularly true if words without pictures are presented with "oos" and "sou" during the learning task); (2) the *gender*-SEX connections account for the difference between the fast and slow learners' ratings, namely that fast learners generalize to new items, but slow learners do not; (3) the *lexical*-*gender* and *gender*-SEX connections are the smallest sufficient set of connections (in addition to the apriori connections) for fitting the model to the human data; hence, in the interest of parsimony no other connections should be added unless they increase the model's explanatory value; (4) the *lexical*-*gender* connections are necessary if CONCEPT-*gender* connections are absent, i.e., the inanimate CONCEPT-SEX connections plus the *gender*-SEX connections cannot guarantee the correct categorization for arbitrary thresholds $ET < CT \in [0, 1]$. For the activation of the *oosative* or *soupative* nodes (for the respective nouns) in networks without *lexical*-*gender* and CONCEPT-*gender* connections are smaller than in networks with *lexical*-*gender* connections and/or CONCEPT-*gender* connections. Hence, it is always possible to choose a threshold value for CT such that networks without *lexical*-*gender* and CONCEPT-*gender* connections incorrectly categorizes all items, while networks with the connections correctly categorizes all items (e.g., by stopping the training as soon as all items have been categorized correctly). In sum, either the *lexical*-*gender* and *gender*-SEX sets of connections or the CONCEPT-*gender* and *gender*-SEX sets are necessary for correct categorization.

Conclusion

The results of the model simulations demonstrate that the consistency effects in both the current and previous rating experiments (e.g., Boroditsky, Schmidt, & Phillips, 2003; Phillips & Boroditsky, 2003), which appear to be due to direct associative links between inanimate concepts and the conceptual properties of sex, can instead be due to indirect associative links between lexical representations and grammatical features and between grammatical features and correlated conceptual properties of sex. More specifically, the results suggest that the absence of any obvious common conceptual property between a picture of a spoon and a picture of a bride leads speakers to base their similarity rating on a common grammatical gender feature of the pictures' names. Although the model does not simulate the other tasks employed by Boroditsky, Schmidt, and Phillips (2003), which have shown an apparent generalization of a correlation between conceptual properties and grammatical features to object concepts, it nonetheless suggests that those results may also be due to indirect associative links, which are utilized to meet the idiosyncratic demands of the task. In short, our model simulations demonstrate the importance of using computational models to test empirical results that appear to support Linguistic Relativity.

Acknowledgments

We thank Frances McCoppin and Carole Kennelly for help with data collection and three anonymous reviewers for their insightful comments. The research was partly supported by a Multiyear Collaborative Research Grant awarded to the first two authors from the University of Notre Dame's Institute for Scholarship in the Liberal Arts.

References

Boroditsky, L., Schmidt, L., & Phillips, W. (2003). Sex, syntax and semantics. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in Mind: Advances in the Study of Language and Thought*. Cambridge, MI: MIT Press.

Phillips, W., & Boroditsky, L. (2003). *Can quirks of grammar affect the way you think? Grammatical gender and object concepts.* Paper presented at the 25th Annual Conference of the Cognitive Science Society, Boston, MA.

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel Distributed Processing: Exploring the Microstructures of Cognition* (Vol. Volume 1: Foundations). Cambridge: The MIT Press.

Sera, M. D., 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*(3), 377-397.

Vigliocco, G., Vinson, D. P., & Paganelli, F. (2004). Grammatical gender and meaning. Paper presented at the 26th Annual Conference of the Cognitive Science Society.