Selective impairment of adjective order constraints as overeager abstraction: An elaboration on Kemmerer et al. (2009)

Bram Vandekerckhove[a,]*, Dominiek Sandra[a], Walter Daelemans[a]

[a] Computational Linguistics & Psycholinguistics (CLiPS) Research Center, University of Antwerp, Prinsstraat 13, 2000 Antwerpen, Belgium

E-mail: {firstname}.{lastname}@ua.ac.be

* Corresponding author. Linguistics Department, University of Antwerp, Prinsstraat 13, 2000 Antwerpen, Belgium. Tel.: +32 3 265 4261; fax: +32 3 265 4259. E-mail address: bram.vandekerckhove@ua.ac.be

Abstract

Kemmerer, Tranel, and Zdanczyk (2009) reported patients who failed to discriminate between preferred and dispreferred orders of prenominal adjectives, yet were sensitive to the order of adjectives in relation to other parts of speech, and able to judge which of two adjectives was most similar to a cue adjective. The authors concluded that knowledge of the semantic constraints on prenominal adjective order can be impaired without an impairment of purely syntactic adjective order knowledge, or knowledge of semantic adjective classes. Using simulation studies, we demonstrate that the impairment of these patients can be characterized as overeager abstraction. Oversmoothing a similarity-based bigram language model with a similarity metric based on word co-occurrence distributions resulted in the same performance dissociation between tasks as reported for Kemmerer et al.'s selectively impaired patients. Additionally, the strength with which the patients preferred a specific adjective order for a given stimulus was predicted by the stimulus' robustness to overeager abstraction. Our results provide a general cognitive account based on the online creation of temporary summary representations that is supported by current neurocognitive views on verbal cognition. This account lends a more insightful explanation for impairments of linguistic knowledge than an explanation relying solely on linguistic abstractions.

Keywords

adjectives; case-based reasoning; representation; symbolic computational modeling; syntax; semantics

In English, there are constraints on how prenominal sequences of descriptive adjectives are ordered. One adjective order is often clearly preferred over the other, e.g., *a large brown desk* vs. *a brown large desk*. These constraints can be associated with the semantic classes of the adjectives involved: the order of prenominal adjectives usually follows a linear precedence relation between their classes. Dixon (1982), for instance, proposed the following class ordering scheme to account for adjective order in a number of languages: VALUE ≺ DIMENSION ≺ PHYSICAL PROPERTY ≺ SPEED ≺ HUMAN PROPENSITY ≺ AGE ≺ COLOR. In two self-paced reading experiments, Kennison (2010) showed that participants' reading times increased at the second adjective in adjective pairs that did not follow this order.

A more generalizing approach to explaining adjective order preferences is to appeal to abstract semantic variables, such as *degree of absoluteness* (the less an attributive adjective's meaning shifts depending on the noun with which it co-occurs, the closer to that noun it tends to be positioned; the interpretation of *brown* is more stable over noun contexts than that of *large*, e.g., *a large brown desk* vs. *a large brown peanut*) (Kemmerer, Tranel, & Zdanczyk, 2009; Martin, 1969; Posner, 1986),[1] or *degree of objectivity* (adjectives denoting relatively inherent, verifiable properties are placed nearer to the noun than adjectives expressing judgments or opinions) (Whorf, 1945; Hetzron, 1978; Quirk, Greenbaum, Leech, & Svartvik, 1985).[2]

Kemmerer et al. (2009) reported patients with brain lesions who showed selectively impaired knowledge of the constraints on prenominal adjective order. They failed a two-alternative forced-choice (2AFC) test that required them to

[1] Posner (1986) referred to this principle as *context independence*. Martin (1969) called it *definiteness of denotation*.

[2] See Wulff (2003) for an analysis of the different factors proposed in the literature.

discriminate between preferred and dispreferred adjective orderings (e.g., *a big brown dog* vs. *a brown big dog*) (Task 1). The same patients were still able to discriminate between correct and incorrect orderings of adjectives in relation to other parts of speech in a noun phrase (e.g., *a cool light rain* vs. *rain light cool a*) (Task 2), and passed a semantic similarity judgment task in which they had to choose which of two adjectives was most similar to a target adjective (e.g., whether the adjective *good* was more similar to *bad* or to *tiny*) (Task 3).

Kemmerer et al. (2009) framed their findings in a construction-based account (Croft, 2001; Goldberg, 1995, 2006; Langacker, 1987, 2008) of prenominal adjective order (see Figure 1 on p. 93 of their paper). Construction-based approaches to language explain linguistic knowledge as form-meaning taxonomies. Kemmerer et al.'s account integrates a class precedence account with principles relying on abstract semantic variables. The form (i.e., syntactic) part of the adjective order construction proposed by Kemmerer et al. stipulates that multiple adjectives can precede a noun. The semantic part specifies that those adjectives denote properties of the noun referent. Semantic constraints (i.e., adjectives denoting objective/absolute properties occur closer to the noun than adjectives denoting subjective/relativistic properties) interact with semantic class-level features of the adjectives (VALUE, SIZE, COLOR, etc.) (meaning) to determine the order of the adjectives (form). In this account, the COLOR adjective *brown* is placed after the SIZE adjective *large,* because COLOR properties are more objective/absolute than SIZE properties.

According to Kemmerer et al. (2009), their findings demonstrated that knowledge of the semantic constraints on prenominal adjective order can be impaired (as shown by the patients' impaired performance on Task 1) without an impairment of knowledge about the purely syntactic constraints on the positioning of adjectives (as shown by the patients' unimpaired performance on Task 2), and with intact knowledge of the adjectives' semantic classes (as shown by the patients' unimpaired performance on Task 3). Kemmerer et al.'s account of their findings is heavily reliant on the theoretical framework they employed, however. In particular, the interpretation that the selectively impaired patients had intact knowledge of the adjectives' semantic classes and must therefore have been impaired in their knowledge of the semantic principles that link those classes with the adjectives' surface order hinges on two assumptions: (a) knowledge of the constraints on adjective order relies on access to persistent abstract representations (features) that correspond to linguistic semantic classes, and (b) a semantic similarity judgment task such as Task 3 of Kemmerer et al. tests for knowledge of those semantic classes. In this paper, we present an alternative account that does not rely on stored summary representations at the level of semantic classes (Tasks 1 and 3) or parts of speech (Task 2) but provides a more fundamental characterization of the selectively impaired patients' behavioral profile as resulting from overeager online abstraction.

Language processing is a prime example of abstract cognitive processing. It seems implausible that the participants of Kemmerer et al. (2009) could exclusively rely on their recollection of the specific word sequences used as stimuli in Tasks 1 and 2 to successfully solve those tasks, as some of the sequences were relatively unusual (e.g., *a terrible narrow bridge*). Instead, they had to have access to mental representations that were more abstract than the specific words in the stimuli. The assumptions on which the account of Kemmerer et al. is built exemplify the common presupposition that abstract mental representations correspond to persistent, stored summary representations—in this particular case, semantic classes for Tasks 1 and 3, and parts of speech for Task 2. However, abstract cognitive processing can also be achieved through online summarization of more concrete representations (Barsalou, 2005). Exemplar models (e.g., Hintzman, 1986; Nosofsky, 1986) store only concrete instances of experience and abstract by creating temporary similarity-weighted summaries of those instances online. This idea of online abstraction over concrete exemplars provides the key to a more insightful, general cognitive account of Kemmerer et al.'s findings that fits in with current neurocognitive views on verbal cognition (e.g., Jefferies & Lambon Ralph, 2006).

Building on the possibility that the participants of Kemmerer et al. (2009) did not refer to stored summary representations but instead relied on online abstraction, we hypothesize that the behavioral pattern of the selectively impaired patients can be described as *overeager abstraction*: while performing the word order tasks (Task 1 and 2), which required the participants to abstract away from the input words for successful performance, the selectively impaired patients activated word representations that were too dissimilar from the input words to still be informative. Instead, those remote representations provided noise that interfered with the formation of task-appropriate summary representations and therefore negatively affected the patients' decisions. According to the overeager abstraction hypothesis, the selectively impaired patients' performance dissociation between Task 1 and Task 2 is explained by the fact that in the former task, the differences between conditions appeal to more fine-grained, less abstract knowledge than in the latter task. In other words, the stimuli of Task 2 are more robust to overeager abstraction than those of Task 1.

Rejecting the assumption that knowledge of the constraints on adjective order relies on access to persistent semantic classes entails rejecting the assumption that Task 3 of Kemmerer et al. (2009) tests for knowledge of those classes. Otherwise, one has to accept the theoretical inconsistency that the participants did not rely on persistent semantic class representations for Task 1, but that they did for Task 3. Because Kemmerer et al. supposed that both Task 1 and Task 3 plug into semantic class knowledge, and because we propose online abstraction as an alternative to predetermined semantic classes, one might then wonder why overeager abstraction should only have a detrimental effect on the former task and not on the latter. However, we claim that the selectively impaired participants of Kemmerer et al. largely solved Task 3 without abstracting away from the stimulus adjectives. Instead, they relied on the strong associations between the cue and target adjectives of Task 3 and the comparatively weaker associations between the cue and distractor adjectives to successfully solve this task.

To test the overeager abstraction hypothesis, we operationalized the idea of online abstraction in a computational implementation of the exemplar model, namely memory-based learning (also known as instance/case-based reasoning/learning) (Fix & Hodges, 1951; Cover & Hart, 1967; Stanfill & Waltz, 1986; Cost & Salzberg, 1993; see Daelemans & van den Bosch (2005) for the application of memory-based learning to linguistic processing). To investigate the hypothesis that Task 3 can be solved on the basis of the prepotent cue-target associations, we compared the performance of a model that used a measure of the adjectives' association strength to that of a similarity-based model.

The computational algorithms are presented in the Section 1. In Section 2, we present two empirical assessments of our hypotheses. In Section 2.1, we report proof-of-concept simulations with the algorithms described in Section 1, which confirm that overeager abstraction can account for the impaired patients' performance dissociation between Tasks 1 and 2, and that an association-based approach performs at least as well on Task 3 as a similarity-based (abstraction-based) approach, without being vulnerable to overeager abstraction. Through a reanalysis of Kemmerer et al.'s (2009) data in Section 2.2, we show that overeager abstraction also accounts for the dissociation in Task 1 performance between the selectively impaired and the unimpaired participants. In Section 3, we discuss the linguistic significance of our results and relate the overeager abstraction hypothesis to Kemmerer et al.'s neuroanatomical findings.

1. Implementing online abstraction and association strength

1.1 A bigram language model

To simulate potential effects of overeager abstraction on Task 1 and Task 2 performance, a computational model needs to (a) be capable of assigning different scores to differently ordered word sequences (preferred/correct vs. dispreferred/incorrect), as word order is what distinguishes the two conditions of Tasks 1 and 2, and (b) contain a parameter that governs online abstraction, i.e., determines how far the temporary summary representations used by the model abstract away from the input. A bigram language model with similarity-based smoothing meets both those requirements.

A word *n*-gram is a sequence of *n* words, with *n* being 2 for a bigram. In its most basic form, an *n*-gram language model approximates the probability of a word sequence $w_1,\ldots,w_\mathrm{m}$ as the product of the conditional probabilities of each word (the *conditioned* word w_i) given the $n-1$ previous words (the *conditioning* words $w_{i-(n-1)},\ldots,w_{i-1}$):

$$P\left(w_1,\ldots,w_m\right) \approx \prod_{i=1}^{m} P\left(w_i \middle| w_{i-(n-1)},\ldots,w_{i-1}\right) \tag{1}$$

The probabilities are derived from frequency counts in a training corpus. The decision to use an *n*-gram language model for scoring the word sequences of Tasks 1 and 2 had two motivations. First, *n*-gram models are close to the simplest possible model for assigning probabilities to word sequences, but have been used extensively and very successfully in natural language processing (Jurafsky & Martin, 2009). Secondly, there is a large body of psycholinguistic evidence that shows the importance of probabilistic predictions about properties of upcoming items on the basis of previous (linguistic) context in human language comprehension. In a visual context, listeners can predict post-verbal arguments on the basis of the verb and its already perceived arguments (Kamide, Altmann, & Haywood, 2003), readers predict the presence of a coordination structure upon reading *either* (Staub & Clifton, 2006), and there are numerous electrophysiological studies showing that sentence processing involves predictions about the semantic and syntactic properties of upcoming material (for a review, see Federmeier, 2007). Because an *n*-gram model can also be seen as a word prediction model, it should be able to capture that aspect of human language processing, especially for the simple word sequences of Tasks 1 and 2.

The two conditions of both tasks consist of the same words and only differ in their word order; so to discriminate between conditions, the minimal size of the *n*-grams should be two. This means that the probability of each word in the sequence is conditioned on the previous word. In an eye-movement study, McDonald and Shillcock (2003) provided evidence that bigram probabilities influence fixation durations, which means that it makes sense to rely on a bigram model for simulating linguistic processing.

1.2 A similarity-smoothed bigram language model

The basic language modeling approach described in the previous section is limited, because it cannot account for people's intuitions about the order of prenominal adjectives for adjectives they have never seen together or in co-occurrence with the head noun before. As discussed in the Introduction, to solve Tasks 1 and 2, the participants of Kemmerer et al.'s (2009) study needed to access information that was more abstract than the stimulus words. Likewise, the bigram language model needs to abstract away from the specific words in the word sequence to cope with bigrams that are not attested in its training input. Backing off to unigram probabilities, which is the usual method to cope with unattested bigrams in language modeling, is not a viable option for adjective order problems. All else being equal, if both $P(Adj_a|Adj_b)$ and $P(Adj_b|Adj_a)$ are zero, backing off to the unigram frequencies of Adj_a and Adj_b would favor the order in which the most frequent adjective appears in the second position. However, in adjective order sequences, there is a tendency for the most frequent adjective to come first (see e.g., Wulff, 2003), which means the model would choose the dispreferred order over the preferred order.

The construction-based account of Kemmerer et al. depends on categorical features such as semantic classes (Task 1) and parts of speech (Task 2) to achieve abstraction. In our account, abstraction is realized online through similarity-based smoothing (see Dagan, Pereira, & Lee, 1994; Dagan, Lee, & Pereira, 1999). This is where language modeling meets memory-based learning (Zavrel & Daelemans, 1997). Memory-based learning (MBL) models are classification models that predict the category of a new observation from stored representations of earlier observations. Their knowledge base is not abstracted from the training data, but consists of a memory of exemplars. Exemplars are usually represented as feature-value vectors, each with an associated category label. During processing, a distance function (metric) is used to assemble the nearest neighbor set, which is the set of memory exemplars at the *k* nearest distances from the test exemplar (because all exemplars at the same distance from the test exemplar are included in the nearest neighbor set). A decision function then determines the test exemplar's category, based on the distribution of categories in the nearest neighbor set (possibly weighted by the distances between the exemplars in the nearest neighbor set and the test exemplar).

The *neighborhood size* (*k*) determines how stringent the similarity requirements for membership of the nearest neighbor set are. At a neighborhood size of 1, the nearest neighbor set is restricted to one nearest neighbor, i.e., the memory exemplar that is most similar to the test exemplar, or all exemplars at that distance, in case of a tie. By increasing the neighborhood size, less similar exemplars are allowed in the nearest neighbor set. Thus, nearest neighbor sets function as summary representations, i.e., they summarize the properties of the exemplars they contain. With the neighborhood size parameter, the memory-based learning model implements a way of varying how abstract these summary representations are. The larger the set, the less properties the members of that set share with the target exemplar, and the more abstract it will be as a summary representation.

The unsmoothed language model described in Section 1.1 can be recast as a memory-based learning model. The feature-value vectors of the exemplars in this model consist of only one variable, namely the conditioning word (w_{i-1}), while the category label is the conditioned word (w_i). Recast as a memory-based learning model, an unsmoothed language model has its neighborhood size set to a value of 1. Similarity-based smoothing consists of increasing the neighborhood size.

In most linguistic classification tasks, e.g. part-of-speech tagging, the categories are representations at a higher level of abstraction than the exemplars and are therefore greatly outnumbered by the exemplar types. Neighbor sets are therefore usually created on the basis of the feature-value exemplar representations. However, in bigram word prediction, the number of categories is usually the same as the number of exemplars, namely the type count in the training corpus,[3] because the set of categories (the conditioned words) is generally the same as the set of exemplars (the conditioning words).[4] In this context, it is therefore as justifiable to smooth over the former as it is to smooth over the latter. The similarity-smoothed language model that was used for the simulations in this study employs the former approach.[5] A model that employs similarity-based smoothing over conditioned words does not estimate conditional probabilities of specific words, but of the nearest neighbor sets of those words. In more formal terms, the model stipulates that the similarity-smoothed probability of a word w_i conditioned on w_{i-1} is given by the summed bigram frequencies of w_{i-1} and the neighbors of w_i at the k nearest distances of w_i ($N_{w_i}^k$) divided by the frequency of w_{i-1} (so that the denominator captures the frequency of all bigrams whose first word is w_{i-1}):

$$P_{SIM}\left(w_i \middle| w_{i-1}\right) = \frac{\sum_{w_i' \in N_{w_i}^k} count\left(w_{i-1}, w_i'\right)}{count\left(w_{i-1}\right)} . \tag{2}$$

When the neighborhood size is set to a value higher than 1, the question of which metric to use becomes relevant. A metric that relies on word co-occurrences

[3] Types include words, punctuation marks, and the sentence delimiter.

[4] That is, if every sentence is preceded and followed by a sentence delimiter.

[5] Dagan et al. (1994) suggested smoothing over conditioned words as an alternative to smoothing over conditioning words. Yet, to our knowledge, there has been no empirical test of this proposal to date.

can be derived directly from the training corpus statistics (Dagan et al., 1999). The Modified Value Difference Metric (MVDM, Stanfill & Waltz, 1986; Cost & Salzberg, 1993) is a metric often used in memory-based learning models, according to which feature values are more similar to the extent that their category distributions overlap. Applied to bigram language models, two words will be more similar as the conditional distributions of the words following them have greater overlap. More formally, the distance between two words w_i and w_j is given by the difference between the distribution of the words conditioned on w_i and the distribution of the words conditioned on w_j,

$$\delta\left(w_i, w_j\right) = \sum_{w_{i+1}} \left| P\left(w_{i+1} \middle| w_i\right) - P\left(w_{i+1} \middle| w_j\right) \right|, \tag{3}$$

where w_{i+1} ranges over all words following w_i or w_j in the training corpus. Identical words have an MVDM of zero.

1.3 The Modified Value Distance Metric, semantic similarity, and word associations

A fundamental idea in computational semantics, first proposed by Harris (1954), is the *distributional hypothesis*, i.e., the hypothesis that words appearing in similar linguistic contexts tend to have similar meanings. This idea has inspired a wide range of distributional semantic models, the best-known being *Latent Semantic Analysis* (LSA) (Landauer & Dumais, 1997) and *Hyperspace Analogue to Language* (HAL) (Lund & Burgess, 1996). As Harris (1954) pointed out, semantic differences between adjectives correlate with differences in their head noun distributions. Because attributively used adjectives are usually immediately followed by their head nouns, MVDM based on the conditional probability distributions of w_{i+1} given w_i should be a good semantic similarity measure for adjectives. In other words, to a great extent, the nearest neighbors of adjectives according to MVDM should be semantically similar adjectives.

Because MVDM with w_{i+1} distributions can be considered a semantic metric, it can be applied to any task that requires judging the semantic similarity between adjectives, such as Kemmerer et al.'s (2009) Task 3. With a small modification to Equation 3, we can calculate the smoothed distance δ_{SM} between nearest neighbor sets instead of the distance between specific words, so as to simulate the effect of online abstraction on the semantic similarity judgments,

$$\delta_{SM}\left(w_i, w_j\right) = \sum_{w_{i+1}} \left| P_{SM}\left(w_{i+1} \middle| w_i\right) - P_{SM}\left(w_{i+1} \middle| w_j\right) \right|, \tag{4}$$

with $P_{SM}(w_{i+1}|w_i)$ given by the conditional probability of w_{i+1} conditioned on the set of neighbors of w_i at the k nearest distances of w_i ($N_{w_i}^{k}$),

$$P_{SM}\left(w_{i+1} \middle| w_i\right) = \frac{\sum_{w_i' \in N_{w_i}^{k}} count\left(w_1', w_{i+1}\right)}{\sum_{w_i' \in N_{w_i}^{k}} count\left(w_i'\right)}, \tag{5}$$

and the distance between w_i and a neighbor w_i' given by MVDM (Equation 3).

Another, simpler approach to Task 3, however, is not to pick the adjective that is most similar to the cue adjective from among the target and distractor adjectives, with similarity defined as the extent to which the adjectives' nearest neighbor sets share post-adjectival context, but to pick the adjective that has the strongest direct word-to-word association with the cue adjective. One simple yet effective way to capture the association strength between words is *pointwise mutual information* (PMI) (Church & Hanks, 1990):

$$\mathrm{PMI}\left(w_i, w_j\right) = \log \frac{P\left(w_i, w_j\right)}{P\left(w_i\right) P\left(w_j\right)}. \tag{6}$$

PMI measures how much the probability of two words w_i and w_j occurring together (their joint probability, $P(w_i,w_j)$, differs from the expected probability of their co-occurrence if they were independent, which is the product of their individual probabilities $P(w_i)P(w_j)$. The higher the PMI of two words is, the more the presence of one word entails the presence of the other, and so the stronger their association. The PMI equation does not contain any parameters to vary online abstraction, which means this method is not affected by overeager abstraction.

2. Testing the overeager abstraction hypothesis

Because the level of abstraction is a parameter in our model, it can be set to a level that is higher than optimal given the task requirements, resulting in overeager abstraction, which we claim to explain the finding of Kemmerer et al. (2009) that the selectively impaired patients failed Task 1 but showed relatively unimpaired performance on Task 2. We hypothesize that the greater robustness to overeager abstraction of the Task 2 items in comparison to the Task 1 items accounts for the dissociation in performance of the selectively impaired patients on those two tasks. This differs with the account of Kemmerer et al., in which the semantic and purely syntactic constraints on adjective order are clearly separable. We also hypothesize that the selectively impaired patients should be more vulnerable to the items' robustness to overeager abstraction than the unimpaired participants within Task 1 itself. More particularly, in the impaired patients' performance, items that are not robust to overeager abstraction should suffer more than items that are robust to a high degree of abstraction. Because the account of Kemmerer et al. relies on abstractions that are predetermined linguistic categories, it predicts no such effect.

A logical concomitant of the overeager abstraction hypothesis is that Task 3 performance, which caused no problems for the impaired participants, did not rely on abstractions, which contrasts with Kemmerer et al.'s (2009) claim that it was based on the use of stable abstract semantic categories (see Introduction). We hypothesize instead that Task 3 can be solved through the use of direct associations between the cue and target adjectives, and that this explains the dissociation in performance of the selectively impaired patients between Task 1 and Task 3.

In the next two sections, we provide evidence supporting the above hypotheses, using the algorithms described in Section 1. The proof-of-concept simulations in Section 2.1 show how the overeager abstraction and direct association hypotheses can account for the dissociations between tasks. The reanalysis of the Task 1 data in Section 2.2 shows how the overeager abstraction hypothesis can account for the dissociation in Task 1 performance between the selectively impaired patients and the unimpaired participants.

2.1 Proof-of-concept: simulating the selective impairment of the semantic constraints on prenominal adjective order constraints

In this section, we investigate the hypotheses (a) that overeager abstraction can explain the selectively impaired patients' dissociation in performance between Tasks 1 and 2 of Kemmerer et al. (2009), and (b) that those patients could solve Task 3 by relying on direct word associations between the stimulus adjectives. In the first two tasks, participants had to choose between differently ordered word sequences. In the third task, they had to indicate which of two adjectives was most similar to a cue adjective. We tested the overeager abstraction and direct association hypotheses by trying to simulate both the behavioral pattern of the selectively impaired patients and that of the unimpaired participants on all three tasks, using the models described in Section 1.

Regarding the effect of overeager abstraction on Task 1 and Task 2 performance, the basic reasoning goes as follows. If a specific word order is supported by the nearest neighbors of the target item words, up to a certain level similarity-based smoothing will confirm or even strengthen the preference for that word order, unless it is exceptional, i.e., very much tied to the specific words involved, such as the atypical order of *big bad* in *the big bad wolf* (which is not in line with the VALUE ≺ SIZE constraint), or *creatures* and *great* in the phrase *all creatures great and small* (which is not in line with the *adjective* ≺ *noun* constraint). Because similarity is based on bigram distributions (see Section 1), the words most similar to the words in the target item should have similar positional preferences. Due to data sparseness, items in the preferred/correct order condition might contain word bigrams that are not attested in the corpus on which the model was trained and so have a zero probability at a neighborhood size of 1, although at the same time bigrams containing very similar words are in effect part of the training data. This means that increasing the neighborhood size should initially have a positive effect on the model's performance for both Task 1 and Task 2, as measured by its accuracy (i.e., the percentage of items for which the model chooses the preferred/correct order). A concrete example, using an item from Kemmerer et al. (2009), makes this clearer. *A nice small cup* has a probability of zero at a neighborhood size of 1, solely due to the fact that *nice small* is not attested in the data on which the model was trained (see Section 2.1.1). The other bigrams, *a nice* and *small cup*, are attested in the training data. However, the bigram *nice large* is attested, and *large* is actually the word that is closest to *small*. This leads to a non-zero probability at a neighborhood size of 2. For the dispreferred order *a small nice cup*, on the other hand, a non-zero probability does not occur until the neighborhood size reaches 30.

However, further increasing the neighborhood size allows words into the neighbor sets that have positional preferences differing considerably from those of the input words, so that noise seeps into the probability distributions, possibly turning the model's choices around or making them equally likely. Because the difference between the conditions of Task 2 involves a higher level of abstraction than the difference between the conditions of Task 1, performance on the former task should be more robust for this noise than performance on the latter task. In other words, the model's performance on Task 1 should start decreasing at a lower neighborhood size than its performance on Task 2. After all, the constraints that are involved in Task 2 can be described at the level of the words' parts of speech. For the similarity-smoothed bigram model to support the correct word order over the reversed order in that task, it basically only matters that the majority of neighbors of the adjectives in the test items are premodifiers and that the majority of neighbors of the test item's noun are nouns. As long as that is the case, the model should assign a higher probability to the correctly ordered noun phrase. It will take a very high level of abstraction before other parts of speech dominate the nearest neighbor set. To solve Task 1, on the other hand, summary representations as abstract as parts of speech are not helpful. More fine-grained distinctions within the category of adjectives need to be made, such as distinctions between (implicit) semantic classes. In order to do so, the model not only needs to be sensitive to the differences in positional preferences between adjectives and nouns, but it should also be sensitive to the differences in positional preferences between adjectives in relation to other adjectives and nouns. Whereas the most similar adjective and noun neighbors will share the positional preferences of the stimulus words to a great extent, once the level of abstraction is increased (i.e., when the neighborhood size is set to a higher value) and neighbors that are less similar to the words in the input are taken into account as well, the model is *oversmoothed*: it becomes blind for the fine-grained differences between the members of the same lexical category, resulting in a performance drop on a task that requires sensitivity to these differences, such as Task 1.[6]

Kemmerer et al. (2009) interpreted the fact that the selectively impaired patients' performance on Task 3 was largely unimpaired as evidence that those patients' semantic class knowledge was still intact. Our account of Kemmerer et al.'s findings replaces static semantic classes with temporary "classes" created by online abstraction. Our model's parallel to intact semantic class knowledge is therefore an intact capability for the online construction of appropriate abstractions. The overeager abstraction hypothesis posits that these patients were impaired in their creation of such abstract representations. Unless one wants to make the claim that those patients' failure to construct the appropriate abstractions during processing was for some reason limited to Task 1 and worked perfectly fine for Task 3, overeager abstraction should also have had an impact on that task. We assessed the effect of overeager abstraction on the semantic similarity judgments by using the smoothed semantic metric (Equation 4) to calculate the distances between the cue and the target/distractor

[6] When the neighborhood size is set to ever higher values, more and more words will become part of the nearest neighbor set of a target word, until eventually the neighbor set includes all words from the training data. At that point, each set of conditioned words is a subset of the nearest neighbor set, and the summed conditional probabilities of these neighbors add up to 1. This results in all sequences having a probability of 1 and the model performing at chance on both tasks.

adjectives, increasing the neighborhood size (*k* parameter) for the calculation of P_{SM} (see Equation 5), and using MVDM with w_{i+1} distributions (Equation 3) to compile the nearest neighbor sets. If smoothed semantic distance is used to solve Task 3, we hypothesize that Task 3 performance will suffer more from a large neighborhood size than Task 2 performance. The reason is the same as why overeager abstraction predicts the dissociation between Task 1 and Task 2: if Task 3 is solved by relying on implicit semantic categories, the distinctions one needs to make for its successful execution are more fine-grained, i.e., less abstract, than the distinctions required for the successful execution of Task 2.

However, we also hypothesize that Task 3 does not test for knowledge of abstract semantic representations, either stored or created online, but can be solved by relying on word-to-word associations. To test this hypothesis, we also used PMI (Equation 6) for the simulation of Task 3. Because calculating the association strength between adjectives in this way does not involve abstraction (there is no *k* parameter in the equation), an overeager abstraction deficit has no bearing on the capacity of the model to tell which of two adjectives is more closely associated with a third one. If a high score on Task 3 is attained using this approach, it puts considerable doubt on the assumption of Kemmerer et al. (2009) that Task 3 tests for abstract semantic knowledge.

2.1.1 Materials and methods

As training material for all algorithms, we used the unannotated 10 million word TASA corpus (Landauer, Foltz, & Laham, 1998). The corpus was lowercased and tokenized before training. The test items were those used by Kemmerer et al. (2009), with only a few minor changes. The reader is referred to that paper for the full list of items. The spelling of a number of items was changed to bring them in line with the training corpus, and in one item a compound noun was replaced by its head noun (*a black metal file cabinet* was changed to *a black metal cabinet*).

For Task 1, the test items consisted of 10 sets of 6 noun phrases composed of an article, followed by two adjectives and a noun, each set combining adjectives from different semantic classes (e.g., *a huge gray elephant* vs. *a gray huge elephant*). Additionally, there were 10 noun phrases with a nominal modifier in the position of the second adjective, the modifying nouns always being material nouns (e.g., *a brown paper bag* vs. *a paper brown bag*). This made 70 test items in total. Kemmerer et al. (2009) embedded these noun phrases in sentences, but because the sentence contexts were identical across conditions, we dropped those for our experiments. The probabilities of the word sequences in Task 1 were approximated by the product of the smoothed conditional bigram word probabilities, with the determiner being the first conditioning word.

The test items of Task 2 comprised 15 items in total. The first five test items consisted of adjective-noun pairs (e.g., *big field* vs. *field big*). The next five items consisted of two adjectives followed by a noun (e.g., *warm sweet air* vs. *air sweet warm*, and the items in the third set consisted of an article followed by two adjectives and a noun (e.g., *a hilly bumpy road* vs. *road bumpy hilly a*). In the items containing two adjectives, those two adjectives were both from the same semantic class, so that according to class-based linguistic accounts such as that of Dixon (1982), there should be no semantic constraints on their order. Because the word sequences started with different words across conditions, the products of the conditional word probabilities were multiplied by the unigram probabilities of the first words.

For Tasks 1 and 2, the similarity-smoothed bigram model chose the sequence containing the most probable order from each word sequence pair using the minimal neighborhood size at which both orders had a non-zero probability, which we will call the *decision neighborhood size*.[7] The model made a random choice if the two orderings were equally probable. In practice, however, this only happened when the neighbor set exhausted the conditional word distribution. Overeager abstraction was simulated by increasing the *lower limit neighborhood size*. As long as the lower limit neighborhood size was lower than the decision neighborhood size, the model used the latter. However, the lower limit neighborhood size replaced the decision neighborhood size if the former was higher than the latter. For these experiments, the lower limit neighborhood size ranged between 1 and 124,793.[8] Because it was not feasible to test all neighborhood sizes, we increased the neighborhood size values at which the model was tested by a factor of 10 at each power of 10. The model scores significantly higher than chance (50%) on Task 1 if it attains an accuracy (percentage of preferred/correct order choices) of 63% (44 out of 70 correct), *p* (two-tailed in a binomial test) = .041.[9] For Task 2, the model scores significantly higher than chance with an accuracy of 80% (12 out of 15 correct), *p* (two-tailed) = .035.

[7] This strategy makes the model more vulnerable to overeager abstraction than if the decision neighborhood size were at the point where at least one of the orders has a non-zero probability. However, earlier experiments showed that the vulnerability to anecdotal evidence of that alternative approach has a more significant negative impact on model performance than the vulnerability to overeager abstraction of our strategy.

[8] 124,793 is the number of types in the training corpus, and so the theoretical upper bound for the neighborhood size parameter.

[9] An α of .05 is used for all statistical hypothesis tests in this paper.

The test items for Task 3 consisted of 25 cue-target-distractor triples (e.g., *wide* vs. *narrow* or *blue*). For the smoothed semantic distance simulation, the model chose the adjective that was closest to the cue, making a random decision if the target and distractor were equally distant from the cue and target. Neighborhood size values increased by a factor of 10 at each power of 10, For the PMI simulation, the model simply chose the adjective that was most strongly associated with the cue, and made a random decision if the PMI values were the same. On this task, performance is significantly higher than chance if the model reaches an accuracy of 72% (18 out of 25 correct), *p* (two-tailed) = .043. For the calculation of PMI, two word tokens were counted as co-occurring if they appeared in the same sentence.

To compile the nearest neighbor sets and calculate the smoothed conditional bigram probabilities, we used the IB1 memory-based learning algorithm as implemented in the Tilburg Memory-Based Learner (TiMBL) machine learning package (Daelemans, Zavrel, van der Sloot, & van den Bosch, 2010), with MVDM as the distance metric, no feature weighting (because the exemplars consisted of only one feature-value pair), random tie resolution and default settings for the other options.[10]

2.1.2 Results and discussion

The results of our simulations with the test items of Kemmerer et al. (2009) are visualized in Figure 1. For Tasks 1 and 2, all items exhausted the set of neighbor candidates at a neighborhood size of 50,000. From that point on, increasing the neighborhood size had no effect on scores anymore for these tasks. For the items of

[10] See the TiMBL manual for more information on these options.

Task 3, the set of neighbor candidates was exhausted at 40,000. All stimulus words of the three tasks were attested in the training data.

INSERT FIGURE 1 ABOUT HERE

Figure 1 shows a marked divergence between the similarity-smoothed bigram model's performance on Task 1 and its performance on Task 2 as the lower limit neighborhood size increases. For Task 1, the model had a mean accuracy of 87% at a lower limit neighborhood size of 1, which is also its peak mean accuracy. Performance stayed more or less stable until the lower limit neighborhood size was 7. After that point, performance on this task more or less steadily decreased, until it reached a low point of 43% at a lower limit neighborhood size of 20,000 and was at chance when the lower limit neighborhood size reached 50,000. The model's performance profile on Task 2 differed noticeably from that on Task 1. The model started off with a mean accuracy of 100% at a lower limit neighborhood size of 1. Although performance dropped to 87% at a lower limit neighborhood size of 200, and more or less stayed at that low until a lower limit neighborhood size of 700, it was back at 100% when the lower limit neighborhood size was at 3,000, and only started declining again from a lower limit neighborhood size of 20,000. Performance on Task 2 reached a low point of 44% at a lower limit neighborhood size of 40,000, and then stabilized at chance level at a lower limit neighborhood size of 50,000.

To check if model performance indeed broke down at smaller values of the lower limit neighborhood size for Task 1 than for Task 2, we compared the items' *breakdown k* values between the two tasks. An item's breakdown *k* is that item's smallest value of the lower limit neighborhood size at which the model assigned either the largest probability to the condition with the dispreferred/incorrect order, or an equal non-zero probability to both conditions. A Wilcoxon rank sum test shows that the breakdown k distributions differed significantly between the two tasks, with breakdown k generally being lower for Task 1 (*Mdn* = 600) than for Task 2 (*Mdn* = 30,000), $W = 205$, $n_1 = 70$, $n_2 = 15$, p (two-tailed) < .001.

The mean accuracy on Task 3 (i.e., the percentage of adjective triads for which the model selected the target adjective as most similar to the cue adjective) of the smoothed semantic distance model (δ_{SM} in Figure 1) started at 76% (it selected the target adjective as most similar to the cue adjective for 19 of the 25 items) for a lower limit neighborhood size of 1. The model reached its peak mean accuracy of 88% (22 out of 25 items correct) at a lower limit neighborhood size of 2. Performance steadily declined with increasing lower limit neighborhood size until the model reached chance performance (50%) at a lower limit neighborhood size of 30,000.

Figure 1 clearly shows that the performance profile of the smoothed distance model on Task 3 is closer to the performance profile of the similarity-smoothed bigram model on Task 1 than to the profile of that latter model on Task 2. This is to be expected if Task 3 is solved on the basis of the online generation of abstract summary representations, as this task (like Task 1) involves more fine-grained distinctions than a task that involves the ordering of syntactic classes (Task 2) and is hence more vulnerable to the addition of distant neighbors. Comparisons of the items' breakdown k values between tasks confirm these profile differences. For Task 3, an item's breakdown k is the smallest value of the lower limit neighborhood size at which the smoothed semantic distance between the cue and distractor adjectives is either higher than or equal to the distance between cue and target. Wilcoxon rank sum tests show there was a significant difference in the distribution of breakdown k between Task 3 (*Mdn* = 800) and Task 2 ($W = 284$, $n_1 = 25$, $n_2 = 15$, p (two-tailed) =

.006), but that the breakdown *k* distributions for Task 3 and Task 1 did not differ significantly, $W = 858$, $n_1 = 25$, $n_2 = 70$, p (two-tailed) = .889.

As an alternative to smoothed semantic distance, we also investigated an approach to Task 3 that compared the PMI of the cue and target adjectives with that of the cue and distractor adjectives. This approach resulted in a mean accuracy of 90%: 22 of the 25 test items were solved correctly, and for one item the model had to make a random decision because target nor distractor appeared in the same sentence as the cue in the training corpus.

Our simulations illustrate the benefits of abstracting away from the input. Because it alleviates the effects of data sparseness, abstraction initially had a positive effect on accuracy for all three tasks. For Task 1, 69 out of 70 items had a decision neighborhood size larger than 1, for Task 2, all 15 items had a decision neighborhood size above 1. Additionally, for Task 3, the similarity-smoothed approach attained its highest mean accuracy of 88% at a lower limit neighborhood size of 2.

However, too much abstraction is eventually detrimental for task performance, but not to the same extent for all tasks. Model performance on Task 1 started to break down at lower neighborhood sizes than performance on Task 2. This supports the overeager abstraction explanation for the selectively impaired patients' dissociation between these two tasks. Overeager abstraction is very harmful for a word order task that requires the language processor to be sensitive to the fine-grained distributional differences between words, such as Task 1. Performance on a word order task that requires the processor to correctly discriminate between the positional preferences of words with different parts of speech, such as Task 2, does not suffer that dramatically from overeager abstraction, because the distributional differences between those words are less subtle. As is shown by our simulations, only when the level of abstraction was so high that the neighbor set taken into account for extrapolation encompassed almost the entire training set of words, performance on this type of task dropped. It seems as if it is almost impossible to have the model break down on this task. Interestingly, this is reminiscent of the finding that canonical word order is actually almost always preserved in aphasia (Bates, Friederici, & Juarez, 1988).

If successfully solving Task 3 requires the online creation of temporary abstract summary representations, overeager abstractors should be more or less equally impaired on that task as on Task 1. This is shown by the closely aligning performance profiles of the similarity-smoothed models on these tasks in Figure 1 and was confirmed by the non-significant outcome of the breakdown *k* comparison. However, our simulation results also illustrate that an approach that does not rely on abstractions but simply on word associations scores at least as well as the similarity-smoothed model at its best neighborhood size setting. This simulation outcome resembles the high performance of the selectively impaired patients in this task, suggesting that those patients did not behave as overeager abstractors when selecting a semantically similar adjective, but relied on word associations instead. This strongly supports our claim that Task 3 does not test for abstract semantic knowledge, neither in the form of stored abstract categories nor in the form of temporary abstractions generated online.

Table 1 shows a systematic comparison between the model and the participant scores. The close correspondence between these scores directly supports the hypothesis that the impaired performance of the patients studied by Kemmerer et al. (2009) on Task 1 was caused by a processing deficit that lead them to overabstract when they extrapolated from stored representations of previous experience, and that the spared performance of the selectively impaired patients on Task 3 was the result of their reliance on associations between largely unimpaired word representations in memory.

INSERT TABLE 1 HERE

Together, the simulations in this section provide empirical support for the hypothesis that overeager abstraction accounts for the dissociation in the selectively impaired patients' performance between Task 1 and Task 2, and for the claim that the relatively unimpaired performance of those patients on Task 3 can be explained by the fact that this task was solved on the basis of word associations in the mental lexicon rather than a comparison of abstract semantic categories (either temporary or permanent ones). The reanalysis of the Task 1 data in the next section tests the hypothesis that overeager abstraction can also account for the performance dissociation between impaired and unimpaired participants on that task.

2.2 Predicting the adjective order choices of the selectively impaired patients

According to Kemmerer et al. (2009), their findings showed that the selectively impaired patients still had knowledge of the adjectives' semantic classes, but were unable to apply that knowledge to the adjective order problem. In their account, abstracting from the input words required access to persistent abstract representations (i.e., the adjectives' semantic class features) and did not involve online summarization of more concrete representations. There is a way to distinguish that explanation from the overeager abstraction hypothesis. If Kemmerer et al.'s view is correct, there should be no systematic relationship between each experimental item's robustness to overeager abstraction, i.e., the neighborhood size at which the model selects the dispreferred adjective order, and the probability that the patients chose the preferred adjective order for that item (i.e., the adjective order that was chosen by the majority of the control group). If, on the other hand, the selectively impaired patients were indeed overeager abstractors, their adjective order preferences should have leaned more towards the preferred adjective order for those items that are more robust to overeager abstraction. Items of the latter kind have a preferred adjective order that is less dependent on the specific words in the noun phrase or their nearest neighbors, but is still supported (i.e., has the highest similarity-smoothed bigram probability of the two response alternatives) when also taking less similar neighbors into account. In other words, in the group of selectively impaired patients, the per-item probability that the preferred adjective order was selected should have been higher for items with a high robustness to overeager abstraction. In the group of unimpaired, i.e., normally abstracting participants, the probability that an item's preferred adjective order was selected should have been influenced less or not at all by how robust the item is to overeager abstraction. We investigated that hypothesis by operationalizing robustness to overeager abstraction as a variable in a reanalysis of the Task 1 data.[11]

2.2.1 Materials and methods

The items of Kemmerer et al.'s (2009) Task 1 differ in the strength with which they lean towards a preferred adjective order, as shown by their *preference biases* (the per-item percentages of preferred order responses) among the control participants. For most items, this preference bias was very high, $M(70) = 0.95$, $SD = 0.086$. Of the 70 test items, 44 had a preference bias of 100%; all 19 control participants chose the

[11] This analysis was necessarily limited to the Task 1 data, because both the selectively impaired and the unimpaired participants' performance on the other two tasks was near or at ceiling.

same adjective order for those items. Twenty-three items had a weaker but still significant preference bias, ranging from 95% ($\chi^2(1, n = 19) = 15.21, p < .001$) to 74% ($\chi^2(1, n = 19) = 4.26, p = .039$). Only for the three items with the smallest preference bias in the control group data did that preference bias not prove significant. These items are *a good big table* and *a yellow plastic toy*, both with 68% agreement on the preferred order ($\chi^2(1, n = 19) = 2.58, p = .108$), and *a long slow train*, with a preference bias of 63% in the control group ($\chi^2(1, n = 19) = 1.32, p = .251$).

To test our hypothesis that an item's robustness to overeager abstraction had a positive effect on the probability that the selectively impaired patients chose the preferred adjective order, we employed mixed logit models with crossed random effects for items and participants (Breslow & Clayton, 1993; Jaeger, 2008), using the R package *lme4* (Bates, Mächler, & Bolker, 2011). The mixed logit model is a type of generalized linear mixed model for binary (i.e., binomially distributed) response variables. Basically, it is an extension of the ordinary logit model that includes both fixed and random effects terms. The decision to use mixed effects models was motivated by the need to account for the non-independence of observations in Task 1, which were grouped within items and participants (see Baayen, Davidson, & Bates, 2008).

Our response variable is *selected adjective order*, with the levels preferred (the order that had the majority of votes in the control group) and dispreferred. The predictors of primary interest are the continuous variable breakdown *k*, which is our operationalization of robustness to overeager abstraction (see Section 2.1.2), and the interaction of breakdown *k* with *participant group* (impaired vs. unimpaired). Additional covariates included in the model are *class distance*, which is the distance between the items' adjective classes on a linear precedence scale, *rating difference*, which is the difference between the naturalness ratings for the preferred and the dispreferred orders, and their interaction with participant group. The inclusion of these covariates is motivated below. The random effects predictors are *item* and *participant*. The model contained the maximal random effects structure justified by the data, based on model comparison. Formally, the log odds of participant *p* selecting the preferred order for item *i,* logit(*SelectedOrder*$_{pi}$ = *Preferred*), is given by the following linear combination of predictors:

$$\begin{aligned}&\text{logit}\left(SelectedOrder_{pi} = Preferred\right) = \\ &\beta_0 + \beta_1 Group\left[Unimpaired\right]_p + \beta_2 RatingDifference_i \\ &+\beta_3 ClassDistance_i + \beta_4 \log(Breakdownk_i) \\ &+\beta_5 RatingDifference_i Group\left[Unimpaired\right]_p \\ &+b_{p0} + b_{i0},\ b_{p0} \sim \mathcal{N}\left(0, \sigma^2_{b_{p0}}\right),\ b_{i0} \sim \mathcal{N}\left(0, \sigma^2_{b_{i0}}\right).\end{aligned} \qquad (7)$$

In the above equation, β_0 is the mean log odds of the preferred adjective order being selected by the selectively impaired patients at the mean values of rating difference, class distance, and log-transformed breakdown *k*; β_1 through β_5 are the parameters for the fixed effects variables. b_{p0} and b_{i0} are the deviations from β_0 for participant *p* and item *i,* respectively (random intercepts). The predictors are summarized in Table 2. In the next paragraphs, we describe them in more detail.

INSERT TABLE 2 ABOUT HERE

The breakdown *k* values were taken from the simulations reported in Section 2. The higher an item's breakdown *k*, the more robust the item is to overeager abstraction. For items with a high breakdown *k*, the model still chooses the preferred order when taking into account neighbor words that are very dissimilar from the words in the test item. For items with a low breakdown *k*, the preferred order is supported only when restricting the neighbors to those words that are most similar to the words in the test item. The breakdown *k* values of the Task 1 items included in the analysis range from 2 to 40,000. As discussed in Section 2, not all neighborhood sizes within that range were tested. Instead they were increased by a factor of 10 at each power of 10. Except for items with a breakdown *k* value not higher than 10, the breakdown *k* values are therefore always upper limit estimates, with the possible difference between an item's actual breakdown *k* and its estimated breakdown *k* being higher for items with high breakdown *k* values. However, most of the Task 1 items included in the analysis have low breakdown *k* values (*Mdn* = 600), so that the deviance between the estimated breakdown *k* and the actual breakdown *k* should be relatively small in the majority of cases. Hence, we do not expect our approach to suffer much from the inaccuracy at higher breakdown *k* values. On the other hand, this predominance of low values for breakdown *k* also means the variable is positively skewed. In order to normalize the data, we used the natural logarithms of breakdown *k* in the model instead of its raw values.

Participant group is a two-level factor with the categories impaired and unimpaired. Kemmerer et al. (2009) isolated six selectively impaired patients, based on the criteria that the patients' scores on Task 1, i.e., their percentage of preferred order responses, had to be more than two standard deviations below the mean score of the control participants, and that the patients had to score more than 90% on Tasks 2 and 3. However, the density plots of the patient and control participant scores for Task 1 in Figure 2 show that the criterion of Kemmerer et al. fails to create the most natural groupings for that task.

INSERT FIGURE 2 ABOUT HERE

Because of the clear dichotomy in the data (see Figure 2) that is not captured by the division Kemmerer et al. (2009) made, and because the majority of patients did not perform any worse than the control group, we decided not to use the groups of

Kemmerer et al. (i.e., 6 selectively impaired patients vs. 19 control participants), but to take all 53 participant scores for Task 1 together (both patients and control participants) and to apply *k*-means clustering with $k = 2$ for the automatic repartitioning of the scores into two groups. *K*-means clustering (MacQueen, 1967) is an unsupervised clustering algorithm that partitions a set of observations into *k* clusters.[12] The standard algorithm starts by randomly selecting *k* observations as initial cluster means, and assigning each observation to the group of the nearest mean. The algorithm then iteratively uses the cluster centroids as new means and reassigns the observations, until it converges. By using this objective method for group assignment, no researcher bias is introduced. As shown in Figure 2, *k*-means clustering does a much better job at identifying those participants that scored exceptionally low on Task 1. Excluding participant 3297, who showed impaired performance on all three tasks and for this reason does not count as selectively impaired, the thus obtained group of impaired patients consists of five participants ($M = 70\%$, $SD = 7\%$), all of whom are among the six patients Kemmerer et al. reported as selectively impaired. One participant that Kemmerer et al. identified as selectively impaired, patient 3273, moves to the unimpaired group, which now consists of 47 participants ($M = 95\%$, $SD = 4\%$). The mean score of the patients in that group ($M = 95\%$, $SD = 4\%$) did not differ significantly from that of the control participants ($M = 95\%$, $SD = 4\%$), $t(42.04) = 0.16$, p (two-tailed) = .872.

Note that the participant group variable is not independent from selected adjective order. On the contrary, the groups are created on the basis of the mean participant scores. This means that the response variable selected adjective order (the

[12] The *k* in the name of the algorithm *k*-means clustering should not be confused with the *k* parameter specifying the neighborhood size in memory-based learning models.

values of which are the participants' responses) contains information on the participant group predictor, so that we expect to find a trivial positive effect of participant group on selected adjective order: the probability that any item is assigned the preferred order will be higher in the group with the highest mean participant scores. However, because of the fact that we are interested in the interaction of breakdown *k* with participant group, we included the participant group variable in the model.

To obtain more reliable estimates for the effects of the main predictors, we also included two covariates and their interactions with the participant group variable in the model. The first covariate, class distance, is derived from the linguistic literature on adjective order. The adjective pairs in the stimuli of Kemmerer et al. are always combinations of two different semantic classes (e.g., PHYSICAL PROPERTY and COLOR for *a soft brown sweater*). As we discussed in the Introduction, these classes interact with adjective order according to a linear precedence relation. Kemmerer et al. used the class precedence scheme VALUE ≺ SIZE ≺ DIMENSION ≺ PHYSICAL PROPERTY ≺ COLOR, which is largely based on the analysis of Bache (1978). Class distance is the number of semantic classes that lies between the classes of the adjectives in the stimulus according to this ordering system. The scheme adopted by Kemmerer et al. uses five classes, which means there are four class distances (0-3). The class distance of *a soft brown sweater* is zero, for instance, because PHYSICAL PROPERTY and COLOR are adjacent on the above scale.

We mentioned in the Introduction that the linear ordering of semantic classes correlates with abstract semantic variables such as degree of objectivity. Hetzron (1978) noted that adjectives at the same level of objectivity are interchangeable (p. 181). From this, it is only a small and arguably uncontroversial step to predict that adjectives that are more distant from each other on these continua will also have stronger ordering preferences (e.g., the more subjective the first adjective is, and the more objective the second, the stronger the preference for adjective 1 $\prec$ adjective 2). Because Kemmerer et al. (2009) provided the semantic classes for all the adjectives in their study, the most straightforward way to approach the underlying continuum is to use those classes.

If semantic classes are good approximations of the level of abstraction used by normal language users to constrain adjective order, the strength with which two prenominal adjectives will tend towards a specific order should correlate positively with their class distance, at least in the unimpaired group. We therefore predict that there will be a positive effect of class distance on selected adjective order in that group, i.e., the higher the class distance, the higher the preference bias for the preferred order. If the selectively impaired patients overabstracted, the online generated temporary representations that determined which adjective order they preferred were more abstract than the temporary representations used by the unimpaired participants, which are well-approximated by semantic classes. This suggests that the distances between semantic classes on the linear precedence scale will be worse predictors of selected adjective order for the selectively impaired participants. Hence, we included an interaction term between participant group and class distance in the model. Class distance was treated as a ratio variable.

The items' class distances correlate strongly enough with their breakdown *k* values, $\tau = 0.35, p < .001$, to raise worries about multicollinearity. As breakdown *k* is the variable of main interest, we want to ensure that it explains variation in selected adjective order that cannot be accounted for by class distance. Therefore, we actually did not use the values of breakdown *k* itself in the model but the residuals of a general linear model with breakdown *k* as the primary variable and class distance as the explanatory variable.

The second covariate included in the model, rating difference, was directly taken from Kemmerer et al. (2009). An item's rating difference is the difference between the average naturalness rating (on a scale from one to five) for the sequence containing the preferred adjective order and the average naturalness rating for the dispreferred sequence. The ratings were provided by 72 college students. Rating difference has no theoretical value in the context of our hypothesis, but because it is a fine-grained measure of order preference strength, it might account for a large part of the variance in selected adjective order that is not explained by the other predictors (as a matter of fact, as a naturalness rating reflects the participants' global assessment of their perceived quality of a particular adjective order, it could be considered as a measure of the impact of all factors that they implicitly take into account when making this assessment). Including it in the analysis could therefore help in obtaining better estimates for the critical predictors, by removing part of the error variance. However, rating difference correlates considerably with both breakdown *k*, $\tau = 0.24$, $p = .01$, and class distance, $\tau = 0.44$, $p < .001$, which could give rise to problems interpreting the contributions of the latter two predictors. Because we do not want rating difference to absorb any variation that the other, theoretically more interesting variables can explain, we did not use the raw values in our model, but the residual errors of rating difference regressed on breakdown *k* and class distance.[13] Of the 70

[13] Note that this is a different problem than the strong correlation between breakdown *k* and class distance, where the unique contribution of breakdown *k* had to be assessed by using its residuals after regressing it on class distance. The difference is that naturalness ratings are global quality assessments reflecting the simultaneous impact

items in the data, naturalness ratings were available for 58 items. Accordingly, that is the number of items included in the analysis. As there were 52 participants, the analysis was performed on 3016 observations.

Kemmerer et al. (2009) found that rating differences correlated with the per-item preference biases, but this correlation was weaker for the seven impaired patients they isolated than for the 19 control participants (pp. 97–98). For that reason, we not only expect to find a positive effect of rating difference on selected adjective order, but also an interaction effect between rating difference and participant group, with the effect of rating difference being stronger in the group of unimpaired participants than in the group of impaired participants.

Before we discuss the results of the regression analysis in the next section, we summarize our hypotheses. Taking the impaired group as the reference level, we expect a positive effect of breakdown *k* on the probability that the selected adjective order is the preferred order, and a negative interaction effect between breakdown *k* and participant group; the effect of breakdown *k* should at least be smaller for the unimpaired participants, although it might still have a positive effect on selected adjective order in that group. We expect to find a (trivial) positive effect of participant group on selected adjective order; the probability that the selected adjective order is the preferred order should be higher in the unimpaired group. Apart from expecting positive effects of class distance and rating difference on selected adjective order in the unimpaired group, we hypothesize that the effect of these predictors will be

of different variables, among which quite likely breakdown *k* and class distance. Accordingly, the variance in rating differences that can be explained by these two variables needs to be removed in order to be able to quantify the independent contribution of these variables.

stronger in that group than in the impaired group, which should manifest itself as positive interaction effects of participant group with class distance and rating difference. We do not have any specific hypotheses as to the question whether class distance and rating difference will have significant effects in the impaired group.

2.2.2 Results and discussion

The results of the mixed logistic regression analysis are summarized in Table 3. We found both a significant positive effect of breakdown *k* on selected adjective order in the impaired group and a significant negative interaction between the effect of breakdown *k* and the effect of participant group, which means that the effect of breakdown *k* on selected adjective order was larger in the impaired group than in the unimpaired group. Additionally, there was a strong, significant effect of participant group, the probability of choosing the preferred order being higher for the unimpaired participants than for the impaired participants. However, as mentioned in the previous section, this is a rather trivial finding, because the response variable (selected adjective order) was used to create the two levels of participant group. We also found that of the two covariates class distance and rating difference, only the former had a significant positive effect on selected adjective order in the group of impaired participants. Additionally, we found significant positive interactions of both these predictors with participant group. This means that both class distance and rating difference had a stronger effect on selected adjective order in the group of unimpaired participants than in the group of impaired participants.

INSERT TABLE 3 ABOUT HERE

We ran a likelihood ratio test to find out whether adding the breakdown *k* predictor to a model that only included the covariates and their interactions with the participant group variable had a significant added value in terms of explanatory power. It had not $(\chi^2(1) = 3.35, p = .068)$, due to the fact that there was no effect of breakdown *k* in the unimpaired group (see below), which is much larger than the impaired group (47 participants vs. 5 participants). However, adding the interaction effect between participant group and breakdown *k* resulted in a model that had significantly more explanatory value than a model without the interaction term ($\chi^2(1) = 4.47, p = .035$). Figure 3 shows the predicted partial effect of breakdown *k* on selected adjective order and its interaction with participant group.

INSERT FIGURE 3 ABOUT HERE

To verify the effects of the main predictor breakdown *k* and the two covariates class distance and rating difference on selected adjective order for the unimpaired participants, we fitted a separate mixed logit model for that group. This analysis showed there was no significant effect of breakdown *k* on selected adjective order, $B = 0.01$ ($SE = 0.05$), $p = 0.799$. On the other hand, selected adjective order was strongly correlated with the items' class distance, $B = 0.95$ ($SE = 0.18$), $p < .001$, and rating difference, $B = 0.99$ ($SE = 0.27$), $p < .001$.

The reanalysis of the Task 1 data confirmed the overeager abstraction hypothesis: the selectively impaired participants were more likely to choose the preferred adjective order for those items with a higher robustness to overeager abstraction, and this behavior distinguished them from the unimpaired participants: only the selectively impaired patients were sensitive to neighborhood interference on their adjective order judgments. Apparently, their restrictions on which word representations were similar enough to the input words to reliably influence their judgments were too lenient. This made them activate mental representations that were too distant from the stimuli and provided support for the adjective order that was dispreferred by the majority of unimpaired participants. For those items with a preferred adjective order that is supported by a large majority of their neighbors, however, this robustness of the positional preferences in the neighbor set made the effect of overeager abstraction less severe, i.e., the impaired participants' decisions were more in line with those of normal, non-overabstracting language users.

Apart from the difference between the selectively impaired and the unimpaired participants concerning the effect of breakdown *k*, we also found that the effect of class distance was smaller for the former participants than for the latter. This finding provides additional support for the overeager abstraction hypothesis. The impaired participants' adjective order preferences were determined by temporary summary representations that were more abstract than the representations normal language users employ. Therefore, the semantic classes that are good predictors of normal language users' adjective order preferences (and which closely correspond to their temporary summary representations, as shown in Section 3.1) were less predictive of the impaired patients' preferences.

Because rating difference and selected adjective order could be said to be both measures of adjective order preference strength (hence their strong correlation in the unimpaired group) it is not surprising that there was only an effect of rating difference in the group of unimpaired participants. If we compare the preference biases (the per-item proportions of preferred order responses) between both groups, we see that there was no significant positive correlation between the preferred order choices of the impaired participants and those of the unimpaired participants, $\tau = 0.11$, *p* (one-tailed) = .159. This suggests that the patients' impairments did not merely result in a less pronounced tendency towards the same preferred adjective orders, but that the adjective orders that were preferred by the impaired participants were often strongly different from the orders preferred by normal language users. It is therefore no surprise that the differences in naturalness ratings provided by normal language users failed to be good predictors of the impaired patients' adjective order choices. Explained in terms of overeager abstraction, the rating differences and the preference biases in the control group are both measures of the adjective order preferences among non-overabstracting language users. Overeager abstraction results in deviating adjective order preferences. This means that differences in naturalness ratings provided by non-overabstracting language users are poor predictors of the preference biases of overeager abstractors. If this interpretation is correct, we should find that the differences in naturalness ratings of the two adjective orders provided by overeager abstractors not only are good predictors of the impaired patients' preference biases, but also, and more importantly, that breakdown *k* is a good predictor of these rating differences: the higher breakdown *k*, the larger the rating differences would be. It would be interesting to investigate this hypothesis in future research.

3. General discussion

We provided empirical evidence that the findings which Kemmerer et al. (2009) described as resulting from impaired knowledge of the semantic constraints governing prenominal adjective order can be explained as resulting from an overeager abstraction impairment: the patients in question created temporary summary representations during language processing that were too abstract to be informative for the adjective order task. In a proof-of-concept simulation study, we showed how oversmoothing a similarity-based bigram language model that used a similarity metric based on word distributions resulted in a behavioral pattern that matched that of the selectively impaired patients of Kemmerer et al. The similarity-smoothed model failed dramatically on Task 1 when abstracting too eagerly, i.e., generalizing too aggressively through its similarity metric, while performance on Task 2 was very robust for overeager abstraction. Apart from that, a simple word association measure achieved a score on Task 3 that came close to the participants' score. This demonstrates that, unlike the other tasks and contrary to the assumption of Kemmerer et al., Task 3 does not require reliance on abstract representations, neither in the form of persistent semantic classes stored in long-term memory, nor in the form of temporary class-like clusters created online. Instead, this task can be solved on the basis of associations between the stimulus words in the mental lexicon.

In the reanalysis of Kemmerer et al.'s (2009) Task 1 data, we found that the strength with which the selectively impaired patients preferred a specific order for two prenominal adjectives in a noun phrase was predicted by that noun phrase's robustness to overeager abstraction (as measured by the noun phrase's breakdown *k*) and that this effect of robustness to overeager abstraction on adjective order preference strength was not shared by the other participants of Kemmerer et al., i.e., the control participants and the non-impaired patients. This is also strong evidence that the selectively impaired patients can be characterized as overeager abstractors. In the remainder of this section, we discuss how our account of Kemmerer et al.'s findings relates to theirs and other linguistic accounts of the adjective order constraints (Section 3.1), and how the account fits the neuroanatomical data (Section 3.2).

3.1 Overeager abstraction and linguistic accounts of adjective order preferences

The bigram language model with similarity-based smoothing we presented in this paper is a general statistical learning model without hard-coded linguistic rules.

However, to the extent that the model implements the abstract linguistic principles that correlate with adjective order (see Introduction), and that the nearest neighbor sets of adjectives can be summarized as semantic classes, it can be seen as *incorporating* a computationally explicit implementation of the construction approach to adjective order that was proposed by Kemmerer et al. (2009). However, the crucial difference between our model and Kemmerer et al.'s adjective order account is the way participants are hypothesized to abstract away from the input. In the model of Kemmerer et al., abstraction requires access to persistent abstract classes that are features of the individual adjectives. In effect, their adjective order construction neatly distinguishes three groups of separable adjective features: (a) the part of speech feature (Adj), which interacts with syntactic constraints to determine the placement of the adjective among other parts of speech in the noun phrase, (b) syntactically relevant categorical semantic features (VALUE, SIZE, COLOR, etc.), which interact with abstract semantic principles to determine the place of the adjective vis-à-vis other adjectives, and (c) syntactically irrelevant word-specific semantic features (GOOD, BIG, RED, etc.). According to Kemmerer et al., knowledge of the abstract part of speech features is required for Task 2, and knowledge of the abstract semantic category features for Tasks 1 and 3. In contrast, our model makes no qualitative distinction between lexical items, semantic classes, or parts of speech; these are just implemented as increasingly larger and therefore increasingly abstract clusters of exemplars. The model posits a single online abstraction mechanism that takes word-level features as input and creates task-appropriate temporary classes that explain adjective order preferences at all three of the lexical, semantic, and syntactic levels; with the caveat that the temporary classes formed by our model's neighbor sets do not necessarily match the linguistic categories (but see below). So whereas Kemmerer et al.'s construction cannot account for "exceptions" such as *the big bad wolf* or *all creatures great and small*, in our account these phrases are fully licensed by the "grammar"; they just require minimal abstraction for correct generalization. Thus, our model is more akin to the construction grammar conceptualization of grammar as a *continuum* of constructions that are linked through a network of constructional schemas (Langacker, 2005) than the adjective order construction that Kemmerer et al. proposed. Translated into construction grammar terms, our explanation for the selectively impaired patients failing Task 1 would be that they employed high-level schemas where more concrete schemas would have been more appropriate.

As opposed to Kemmerer et al.'s (2009) account, our model posits that Task 1 and Task 2 are not processed differently because of a qualitative difference in the order violation (syntactic vs. semantic constraints), but because of a quantitative difference. In our model, constraints on word order are on a continuum from lexical constraints over grammatical-semantic constraints to purely syntactic constraints, which is reflected in the level of abstraction that is necessary for correct generalization. In the first case, the so-called "syntactically irrelevant" (Kemmerer et al., 2009, p. 93) lexical features turn out to be very important, because even the slightest abstraction beyond the specific input words will make the processor prefer the alternate order. This is reflected in the breakdown *k* values. For example, because *the big bad wolf* has an idiosyncratic order very much tied to the specific words involved —VALUE adjectives usually precede SIZE adjectives—the memory-based model can correctly distinguish between the preferred and dispreferred orders at a neighborhood size of 4, but already chooses the dispreferred, reversed order at a neighborhood size of 5. The preferred order of this phrase goes against order preferences of bigrams containing near neighbors of *big* such as *bad little* and *bad* *old*. In other words, the preferred order at the lexical level is no longer the preferred order at a more abstract level, where SIZE adjectives tend to stand closer to the head noun than VALUE adjectives. In comparison, the phrase *the good little wolf*, containing semantically similar words, but with a non-idiosyncratic order according to the semantic class precedence, has a decision neighborhood size of 7, but a breakdown *k* between 20 and 30. Finally, it takes a neighborhood size between 100 and 200 for the model to lose its "purely syntactic" (Kemmerer et al., 2009, p. 96) preference for *the big bad wolf* over *wolf big bad the*. In this context, it is worth pointing out that the dissociation between Task 1 and Task 2 of the selectively impaired patients was not that clear-cut at all (a fact acknowledged by Kemmerer et al.). Of the five patients that were automatically categorized as selectively impaired for the purpose of this paper (see Section 2.2.1), two patients still made a mistake in Task 2, which contrasts with the perfect scores achieved by all control participants. This is hard to explain in a model that firmly distinguishes between types of representational features, such as the adjective order construction of Kemmerer et al. Indeed, if a particular type of representational feature is left intact, a task requiring the use of this feature should give rise to error-free performance. In contrast, this finding is very straightforward to explain in a model that acknowledges a continuum between representational levels. This point is nicely illustrated by the performance profile in Figure 1.

Focusing on adjectives, a bigram language model implements the linguistic principles relying on abstract semantic variables that we discussed briefly in the Introduction and whose selective impairment is the source of the patients' adjective ordering problems according to Kemmerer et al. (2009). The key to understand this link is the fact that those linguistic principles can at least partially be derived from correlations between the order of the adjectives and the differences between their respective adjective-noun co-occurrence distributions. In case of the semantic principle that relates the adjectives' degrees of objectivity to their order, the link with the bigram language modeling approach is straightforward. Subjective adjectives have wider *ranges of applicability* than objective adjectives.[14] The former can be applied to a larger set of nouns, because their felicitousness depends more on the opinion of the speaker than on their fit with the head noun. Adjectives denoting inherent, objective properties of certain noun referents will occur more often with those specific nouns and less often with other nouns. If an adjective can occur with a wide range of nouns, the total probability mass of $P(noun|adjective)$ will be distributed over more nouns, which lowers the probability of any single noun co-occurring with the adjective. If an adjective co-occurs with a limited number of nouns, the probability that it occurs with any of those nouns will be relatively high. All other things being equal, this means that the bigram-approximated probability of a felicitous sequence containing an objective adjective immediately followed by a noun will in general be higher than that of a sequence containing a subjective adjective immediately followed by the same noun. Hence, the linguistic principle that objective adjectives occur closer to the noun automatically follows from the model.

Other semantic principles put forward in the linguistic literature on adjective order can also be related to distributional facts. Although a semantic variable such as degree of absoluteness (see Introduction), for instance, is in itself not a distributional measure, this variable is indirectly related to adjective-noun co-occurrence distributions. After all, an adjective that can co-occur with a large range of nouns will

[14] Range of applicability itself has been proposed as a principle underlying prenominal adjective order by Ziff (1960), who called it *privilege of occurrence*, and Seiler (1978).

be less specific in its meaning and more flexible in its interpretations than an adjective that can only be used felicitously with a restricted set of nouns. Hence, the model predicts higher co-occurrence probabilities with the noun for absolute adjectives, which is in line with the linguistic principle regarding absoluteness.

Likewise, the fact that MVDM with w_{i+1} distributions can be considered a measure of adjective similarity (see Section 1.3) relates this metric to the adjective order model of Kemmerer et al. (2009) and theoretical-linguistic accounts of adjective order in general. First, because MVDM is a distributional measure, the nearest neighbors of adjectives given MVDM with w_{i+1} distributions will have similar ranges of applicability. This means that the metric respects the linguistic principles that rely on abstract semantic variables such as degree of objectivity and degree of absoluteness, which can be framed in terms of distributional properties, as we discussed earlier. Secondly, because the nearest neighbors of adjectives according to MVDM with w_{i+1} distributions are semantically similar adjectives, the nearest neighbor sets created with this metric should at least partially overlap with sets formed on the basis of the adjectives' semantic classes, which are adjective order determinants in class-based accounts such as that of Kemmerer et al.

What sets our approach apart from class precedence accounts of adjective order preferences such as the one by Kemmerer et al. (2009) is that the "classes" formed by the nearest neighbor sets are not retrieved as persistent abstract representations, but are created on the spot as a result of online abstraction. The account we have introduced in this paper hinges on the idea that the processing of noun phrases involves the ad hoc creation of implicit, temporary summary representations over which conditional probabilities are computed. These temporary "classes" can be approximated by sets of words that are similar to the target words according to a distributional similarity metric. The level of abstraction of these implicit classes is a task-dependent variable, in our memory-based learning approach implemented as the neighborhood size parameter *k*. The more lenient the similarity requirements for class membership are, the more abstract the implicit classes (eventually "coinciding" with syntactic classes). Due to the fact that these word classes are not persistent cognitive summary representations, but are created online, cognitive deficits can interfere with this online abstraction process. One way in which the temporary class creation might go wrong is when the similarity requirements for class membership are too lenient, resulting in overeager abstraction.

As illustrated in Table 4 for the color adjectives that Kemmerer et al. (2009) used in their stimuli, the nearest neighbor sets that result from employing the similarity metric at a fairly low level of abstraction often correspond to or form subclasses of the semantic classes that are used as explanatory constructs in some adjective order theories. Hence, our observation that the distances between those semantic classes on a linear precedence scale (class distance) correlate so strongly with the preferred adjective order biases of Kemmerer et al.'s unimpaired participants (see Section 2.2.2) can be explained in terms of temporary "semantic classes" that are generated online. Indeed, as we showed in our reanalysis of their Task 1 data, these findings do not mean that impaired access to stable semantic classes accounts for the behavior of the selectively impaired patients. The stimuli's robustness to overeager abstraction was a significant predictor of the items' preferred adjective order biases in the selectively impaired group, and in the same participant group class distance was a worse predictor of these preferred order biases. The relation between the strength with which the choices of the selectively impaired patients tended towards a specific adjective order and the level of abstraction it took before our model lost its preference for that order (breakdown *k*) suggests an online process of abstraction to arrive at temporary summary representations, instead of a process that makes use of persistent abstract representations corresponding to the semantic classes from theoretical linguistics. Hence, while our account is in line with the linguistic principles behind adjective order, it rejects the assumption that these principles operate on fixed abstract representations of semantic classes.

INSERT TABLE 4 ABOUT HERE

Kemmerer et al. (2009) included Task 3 as a test for the participants' semantic class knowledge. In Section 2.1, however, we saw that simply applying a measure of word association strength to the task without abstracting away from the adjective stimuli already resulted in an accuracy of 90%. This finding seriously puts into doubt that Task 3 tested for the participants' knowledge of (implicit) semantic classes. On the basis of their Task 3 findings, Kemmerer et al. proposed that the selectively impaired patients still had knowledge of semantic adjective classes, but that somehow those patients could not use that knowledge to guide their adjective order decisions, as opposed to normal language users. If we apply the same reasoning to our account, we would have to make the contradictory claim that online abstraction failed for the selectively impaired patients when they had to judge adjective order, as they could not appropriately create abstract temporary summary representations, but that this creation of summary representations worked just fine when they had to judge the similarity between adjectives. A more straightforward explanation for the performance dissociation between Task 1 and Task 3, one that is in line with our model, is that the selectively impaired patients had an impairment that caused them to overabstract when a task required abstraction from the input, but that this impairment was less of a problem for a task that did not require abstraction and could largely be solved through associations between the cue and target adjectives that are stored in the mental lexicon. This account is largely compatible with current neurocognitive approaches to semantic knowledge recovery (e.g., Jefferies & Lambon Ralph, 2006, Wagner, Paré-Blagoev, Clark, & Poldrack, 2001), in which a distinction is made between top-down controlled access to long-term semantic knowledge and bottom-up association-based access processes. Tasks requiring access to abstract representations (in our account temporary summary representations), such as the word order tasks 1 and 2, might rely more on the former route, while tasks that can be solved through pre-existing associations between words, such as Task 3, might rely more on the latter route. In the next section, we discuss a way to interpret overeager abstraction as a selective neurological impairment of the top-down route of controlled memory retrieval.

3.2 Overeager abstraction as a neuropsychological impairment

Overeager abstraction is essentially an impairment of the controlled access to knowledge stored in memory. It affects the online creation of task-appropriate summary representations, but not the long-term memory representations of the exemplars that provide input to this summarization process. In what follows we will show that this cognitive characterization of the patients' selective impairment is in line with Kemmerer et al.'s (2009) neuroanatomical findings.

Kemmerer et al. (2009) found that among the patients with unilateral left hemisphere damage, the most frequently affected regions were the posterior inferior frontal gyrus and the underlying white matter, and the inferior parietal lobule.[15] The

[15] Apart from the six selectively impaired patients Kemmerer et al. (2009) identified as selectively impaired, they also included six patients from an earlier study by

left inferior frontal gyrus (LIFG) has been shown to be important for controlled memory retrieval (e.g., Bunge, Wendelken, Badre, & Wagner, 2005; Wagner et al., 2001). Importantly, the region has been implicated in the selection of memory information under competition, or the inhibition of irrelevant alternatives, not only when task demands explicitly require the selection of appropriate responses from among competitors (e.g., Thompson-Schill, D'Esposito, Aguirre, & Farah, 1997; Thompson-Schill et al., 1998; Zhang, Feng, Fox, Gao, & Hai Tan, 2004), but also when selection is implicit (Grindrod, Bilenko, Myers, & Blumstein, 2008).

The overeager abstraction hypothesis is compatible with the proposed role of LIFG in selection and/or inhibition. Our model implements the idea that the word categories determining the language processor's sensitivity to word order during comprehension are temporary products of an online abstraction process that is constrained from the top down by the comprehension goal and already processed context. To create the right categories given the comprehension goal, selection of relevant and suppression of irrelevant information is crucial. When the neighborhood size parameter in our model is set to a value that is too high, weeding out irrelevant information or restraining alternatives, i.e., disregarding irrelevant neighbor words, is exactly what the model fails to do. This is most detrimental for phrases with a low

Kemmerer (2000) in their neuroanatomical analysis. Kemmerer (2000) reported patients with selectively impaired knowledge of adjective order constraints, but did not include a semantic similarity judgment task like Task 3 of Kemmerer et al. (2009). Kemmerer et al. (2009) conceded that the type of impairments of the patients from the earlier study might therefore not overlap completely with those of the patients they reported, but maintained that the deficit both groups of patients shared was functionally still well constrained.

breakdown *k* value, for which word order preferences are strongly tied to the specific words involved. Those cases require the model to be very inhibitive in its selection of relevant neighbors, because even the closest neighbors provide support for the reversed adjective order. Hence, the evidence that the LIFG is important for the suppression of irrelevant information is compatible with our concept of an overeager online abstraction process in Kemmer et al.'s (2009) patients.

As part of a two-component theory of semantic cognition, the temporoparietal region has also been postulated to support the task-dependent executive control of semantic activation, together with the left inferior prefrontal cortex (e.g., Corbett, Jefferies, & Lambon Ralph, 2011; Jefferies & Lambon Ralph, 2006; Noonan, Jefferies, Corbett, & Lambon Ralph, 2010). According to this approach, control processes subserved by these regions regulate the access and use of amodal semantic representations that are stored in the anterior temporal lobes (ATL). Damage of the left inferior prefrontal and/or temporoparietal cortex results in *semantic aphasia* (SA) (Noonan et al., 2010), an impairment that can be characterized as a deregulation of the task-appropriate activation of semantic knowledge. At least within the semantic domain, so far no distinction in behavioral profiles has been found between patients with lesions in one, the other, or both of these regions (Corbett et al, 2011; Jefferies & Lambon Ralph, 2006). Interestingly, in their discussion of the SA behavioral profile, both Jefferies and Lambon Ralph (2006) and Crutch and Warrington (2008) refer to Goldstein's (1948) characterization of aphasia as a loss of "abstract attitude". Although this characterization is exactly the opposite of the overeager abstraction impairment that we claim typifies the selectively impaired patients of Kemmerer et al. (2009), the shared focus on the lacking capacity to adequately abstract away from concrete knowledge of both our and Goldstein's accounts is notable. However, our study broadens this view by showing that a control impairment or impairment of the "abstract attitude" does not need to lead to lazy abstraction, but can just as well lead to the opposite behavior, i.e., overeager abstraction.

There are a number of interesting correspondences between the behavior of semantic aphasics and the selectively impaired patients reported by Kemmerer et al. (2009). The performance of SA patients is typically inconsistent between tasks, dependent on the type of semantic processing required. This is in line with the performance dissociations reported by Kemmerer et al. Additionally, SA performance does not seem to correlate with item frequency, as opposed to the performance of patients with semantic dementia, a semantic memory impairment associated with atrophy of the frontal temporal lobe. In line with this, Kemmerer et al. reported that the impaired patients' performance did not significantly correlate with the means of the first and second adjective frequencies (p. 97). Our claim that the selectively impaired patients could solve Task 3 by reliance on bottom-up associations also fits the SA behavioral profile. SA only affects top-down controlled access to semantic knowledge, not the memory representations themselves or the automatic, bottom-up semantic associations between those representations. Interestingly, Jefferies and Lambon Ralph (2006) found prepotent associations to be a frequent cause of errors in SA patients (e.g., responding *nuts* for *squirrel* in a picture naming task), showing that their sensitivity for automatic, bottom-up semantic processing was still intact. The selectively impaired patients' reliance on automatic word associations for Task 3 is exactly what we claimed resulted in their relatively unimpaired performance on that task (see Section 2.1). Moreover, if Task 3 required the use of semantic category knowledge, the finding that patients with prefrontal and/or temporoparietal lesions were impaired on category-based similarity judgments (Noonan et al., 2010) (e.g., is *broccoli* most similar to *cauliflower* or to *lobster*; the relationship in this case being the category *vegetables*), would lead one to expect the selectively impaired patients of Kemmerer et al. to perform poorly on this task. The fact that this was not the case supports our view that Task 3 did not require such knowledge and could be solved by using simple word associations.

Apart from patient studies that show the importance of the temporoparietal region for controlled knowledge retrieval, there is considerable neuroimaging evidence supporting the role of this region in semantic processing that is particularly relevant in light of the supposed connection between abstract category knowledge and word order. Based on a meta-analysis of 120 fMRI studies, Binder, Desai, Graves, and Conant (2009) see an important high-level role for specifically the angular gyrus in tasks that require the fluent combination and integration of concepts, such as the comprehension of complex noun phrases. More specifically, Cristescu, Devlin, and Nobre (2006) showed that both left inferior frontal and left inferior parietal areas were activated when participants were cued with the semantic categories of upcoming words, and suggested that these regions play a role in the generation of semantic expectations, a notion that is compatible with our distributional account of how Task 1 is solved. The view that the sensitivity to word order during comprehension of adjective order sequences (as in Kemmerer et al.'s (2009) Task 1) relies on the fit between the expected and the actual semantic categories of the adjectives, either explicit (Kemmerer et al., 2009) or implicit (our model), is in line with these neuroimaging findings. Moreover, because the lesion patterns of the selectively impaired patients were primarily in the regions that might be involved in attentional orienting to semantic category information according to Cristecscu et al.'s (2006) study, their findings provide additional support against Kemmerer et al.'s view that the selectively impaired patients with damage in those areas were still capable of directing attention to the semantic categories of the adjectives, and for our view that Task 3 mainly required automatic, association-based memory access.

Interestingly, posterior LIFG has also been shown to be important for syntactic processing and the sequencing of both linguistic and non-linguistic stimuli; Thothathiri, Schwartz, and Thompson-Schill (2010), for instance, reported patients with lesions including the junction of Brodmann areas 44 and 6 who showed erroneous picture-cued production of conjoined noun phrases (e.g., *The eye and the pencil*), particularly when the phrases followed a non-canonical order, i.e. when the first noun was inanimate and the second one animate (e.g., *The shoe and the cat*). The same patients showed impaired comprehension of non-canonical reversible sentences, in which subjects did not correspond to agents, such as passive sentences (e.g., *the man is photographed by the woman)*. The authors suggested that, to fit task goals, the posterior ventrolateral prefrontal cortex might bias activations between representations that are activated in long- and short-term memory on the one hand, and bottom-up input activations on the other. They characterized the impairment of the patients with lesions in those areas as the inability to override recent experience or prevalent patterns in long-term memory. Their account focused on the competition between the information sources, but the findings also support an interpretation in terms of overeager abstraction. The patterns that the patients described by Thothathiri et al. (2010) had trouble overriding, such as ANIMATE ≺ INANIMATE or *subject* = *agent*, are abstract schemas, involving abstract features such as animacy, syntactic functions and thematic roles. In theory, problems with competition between different information sources can just as well consequently favor the more concrete bottom-up information, which would not negatively affect the conjoined noun phrase production task of Thothathiri et al. (2010). The finding that the patients they reported had trouble ignoring the abstract regularities in favor of the concrete order information in the visual input makes those patients in fact overeager abstractors.

As we mentioned before, Kemmerer et al.'s (2009) adjective order construction distinguishes between idiosyncratic features that have no influence on syntax, semantic categories that influence word order among adjectives, and syntactic categories that determine what position the adjectives take vis-à-vis nouns and other parts of speech. In our account, on the other hand, all three levels are syntactically relevant, because order constraints can operate at all levels of abstraction. Thus, online abstraction might actually provide a general framework for unifying the different linguistic functions that are subserved by the LIFG, as we have seen above that it underlies both semantic and syntactic processing. Moreover, the cognitive control function of the (posterior) LIFG extends beyond language (see Novick, Trueswell, & Thompson-Schill, 2010, for a review). We do not have any information on how the selectively impaired patients of Kemmerer et al. performed in non-linguistic tasks. However, there is nothing specifically linguistic about the online abstraction hypothesis, so it might even be relevant outside of the linguistic domain. We plan on future work to assess whether overeager abstraction has validity beyond the linguistic domain and should be considered as a symptom of a general neuropsychological impairment.

Acknowledgments

This research was supported by a grant from the Research Foundation – Flanders (FWO) to the first author. The TASA corpus was made available to us by courtesy of Thomas Landauer, University of Colorado at Boulder. We are very grateful to David

Kemmerer for providing us with the data of Kemmerer et al. (2009). We thank Kevin Diependaele and Peter Goos for helping us with the statistical analyses in Section 3. Jen Corbett and an anonymous reviewer provided invaluable comments on earlier versions of this article.

References

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.

Bache, C. (1978). *The order of premodifying adjectives in present-day English*. Odense: Odense University Press.

Barsalou, L. W. (2005). Abstraction as dynamic interpretation in perceptual symbol systems. In L. Gershkoff-Stowe & D. Rakison (Eds.), *Building object categories* (pp. 389–431). Majwah, NJ: Erlbaum.

Bates, E., Friederici, B. W., & Juarez, L. (1988). On the preservation of word order in aphasia: Cross-linguistic evidence. *Brain and Language, 33*, 323–364.

Bates, D., Mächler, M., & Bolker, B. (2011). lme4: Linear mixed-effects models using S4 classes (R package version 0.999375-42) [Software]. Available from CRAN: http://CRAN.R-project.org/package=lme4

Binder, J. R., Desai, R. H., Graves, W. W., & Conant L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex, 19*, 2767–2796.

Breslow, N.E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association, 88*, 9–25.

Bunge, S. A., Wendelken, C., Badre, D., & Wagner, A. D. (2005). Analogical reasoning and prefrontal cortex: evidence for separable retrieval and integration mechanisms. *Cerebral Cortex, 15*, 239–249.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics, 16*, 22–29.

Corbett, F., Jefferies, E., & Lambon Ralph, M. A. (2011). Deregulated Semantic Cognition Follows Prefrontal and Temporoparietal Damage: Evidence from the Impact of Task Constraint on Nonverbal Object Use. *Journal of Cognitive Neuroscience, 23*, 1125–1135.

Cost, S., & Salzberg, S. (1993). A weighted nearest neighbor algorithm for learning with symbolic features. *Machine Learning, 10*, 57–78.

Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13,* 21–27.

Cristescu, T. C., Devlin, J. T., & Nobre, A. C. (2006). Orienting attention to semantic categories. *NeuroImage, 33*, 1178–1187.

Croft, W. A. (2001). *Radical construction grammar: syntactic theory in typological perspective*. Oxford: Oxford University Press.

Crutch, S. J., & Warrington, E. K. (2008). Contrasting patterns of comprehension for superordinate, basic-level, and subordinate names in semantic dementia and aphasic stroke patients. *Cognitive Neuropsychology, 25*, 582–600.

Daelemans, W., & van den Bosch, A. (2005). *Memory-based language processing*. Cambridge: Cambridge University Press.

Daelemans, W., Zavrel, J., van der Sloot, *K*., & van den Bosch, A. (2010). *TiMBL: Tilburg Memory Based Learner, version 6.3, Reference Guide* (ILK Technical Report No. 10-01). Tilburg: Tilburg University.

Dagan, I., Pereira, F., & Lee, L. (1994). Similarity-based estimation of word co-occurrence probabilities. *Proceedings of the 32nd annual meeting of the Association for Computational Linguistics* (pp. 272–278). Morristown, NJ: Association for Computational Linguistics.

Dagan, I., Lee, L., & Pereira, F. C. N. (1999). Similarity-based models of word co-occurrence probabilities. *Machine Learning, 34*, 43–69.

Dixon, R. M. W. (1982). *Where have all the adjectives gone?* Berlin: Mouton de Gruyter.

Federmeirer, *K*. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology, 44*, 491–505.

Fix, E., & Hodges, J. L. (1951). *Discriminatory analysis—nonparametric discrimination: consistency properties* (Technical Report Project 21-49-004-11, No. 4). Randolph Field, TX: USAF School of Aviation Medicine.

Goldberg, A. (1995). *Constructions: a construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Goldberg, A. (2006). *Constructions at work: the nature of generalization in language*. Oxford: Oxford University Press.

Goldstein, K. (1948). *Language and language disturbances: aphasic symptom complexes and their significance for medicine and theory of language*. New York: Grune & Stratton.

Grindrod, C. M., Bilenko, N. Y., Myers, B. M., & Blumstein, S. E. (2008). The role of the left inferior frontal gyrus in implicit semantic competition and selection: An event-related fMRI study. *Brain Research, 1229*, 167–178.

Harris, Z. (1954). Distributional structure. *Word, 10*, 146–162.

Hetzron, R. (1978). On the relative order of adjectives. In H. J. Seiler (Ed.), *Language universals* (pp. 165–184). Tübingen: Narr.

Hintzman, D. (1986). Schema abstraction in a multiple-trace memory model. *Psychological Review, 93*, 411–428.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*, 434–446.

Jefferies, E., & Lambon Ralph, M. A. (2006). Semantic impairment in stroke aphasia versus semantic dementia: A case-series comparison. *Brain, 129*, 2132–2147.

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, speech recognition, and computational linguistics* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language, 49*, 133–156.

Kemmerer, D. (2000). Selective impairment of knowledge underlying prenominal adjective order: Evidence for the autonomy of grammatical semantics. *Journal of Neurolinguistics, 13*, 57–82.

Kemmerer, D., Tranel, D., & Zdanczyk, C. (2009). Knowledge of the semantic constraints on adjective order can be selectively impaired. *Journal of Neurolinguistics, 22*, 91–108.

Kennison, S. (2010). Processing prenominal adjectives during sentence comprehension. *Perceptual and Motor Skills*, *111*, 141–157.

Landauer, T. K, & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review, 104*, 211–240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25*, 259–284.

Langacker, R. W. (1987). *Foundations of cognitive grammar, vol. I, theoretical prerequisites*. Stanford: Stanford University Press.

Langacker, R. W. (2005). Construction Grammars: cognitive, radical, and less so. In M. S. Peña Cervel & F. J. Ruiz de Mendoza Ibáñez (Eds.) *Cognitive linguistics: internal dynamics and interdisciplinary interaction* (pp. 101–159). Berlin: Mouton de Gruyter.

Langacker, R. W. (2008). *Cognitive Grammar*. Oxford: Oxford University Press.

Lund, K. & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments & Computers, 28*, 203–208.

McDonald, S. A., & Shillcock, R. C. (2003). Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science, 14*, 648–652.

MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. In L. M. Le Cam & J. Neyman (Eds.), *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (pp. 281–297). Berkeley, CA: University of California Press.

Martin, J. E. (1969). Semantic determinants of preferred adjective order. *Journal of Verbal Learning and Verbal Behavior, 8*, 697–704.

Noonan, K. A., Jefferies, E., Corbett, F., & Lambon Ralph, M. A. (2010). Elucidating the Nature of Deregulated Semantic Cognition in Semantic Aphasia: Evidence for the Roles of Prefrontal and Temporoparietal Cortices. *Journal of Cognitive Neuroscience, 22*, 1597–1613.

Nosofsky, R. (1986). Attention, similarity and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39–57.

Novick, J. M., Trueswell, J. C., & Thompson-Schill, S. L. (2010). Broca's area and language processing: Evidence for the cognitive control connection. *Language and Linguistics Compass, 4*, 906–924.

Posner, R. (1986). Iconicity in syntax. In T. A. Sebeok & P Bouissac (Eds.), *Iconicity: essays on the nature of culture: festschrift for Thomas A. Sebeok on his 65th birthday* (pp. 305–337). Tübingen: Stauffenburg.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. London: Longman.

Sarkar, D. (2010). lattice: Lattice Graphics (R package version 0.18-3) [Software]. Available from CRAN: http://CRAN.R-project.org/package=lattice

Seiler, H. (1978). Determination: A functional dimension for interlanguage comparison. In H. Seiler (Ed.), *Language universals* (pp. 301–328). Tübingen: Narr.

Stanfill, C., & Waltz, D. (1986). Toward memory-based reasoning. *Communications of the ACM, 29*, 1213–1228.

Staub, A., & Clifton, C. (2006). Syntactic prediction in language comprehension: Evidence from *either ... or*. *Journal of Experimental Psychology: Learning, Memory and Cognition, 32*, 425–436.

Thompson-Schill S. L., D'Esposito, M., Aguirre, G. K., & Farah, M. J. (1997). Role of left inferior prefrontal cortex in retrieval of semantic knowledge: A reevaluation. *Proceedings of the National Academy of Sciences, 94*, 14792–14797.

Thompson-Schill, S. L., Swick, D., Farah, M. J., D'Esposito, M., Kan, I. P., &

Knight, R. T. (1998). Verb generation in patients with focal frontal lesions: A neuropsychological test of neuroimaging findings. *Proceedings of the National Academy of Sciences, 95,* 15855–15860.

Thothathiri, M., Schwartz, M. F., &Thompson-Schill, S. L. (2010). Selection for position: The role of left ventrolateral prefrontal cortex in sequencing language. *Brain and Language, 113*, 28–38.

Wagner, A. D., Paré-Blagoev, E. J., Clark, C., & Poldrack, R. A. (2001). Recovering meaning: Left prefrontal cortex guides controlled semantic retrieval. *Neuron, 31*, 329–338.

Whorf, B. L. (1945). Grammatical categories. *Language*, *21*, 1–11.

Wulff, S. (2003). A multifactorial corpus analysis of adjective order in English. *International Journal of Corpus Linguistics, 8*, 245–282.

Zavrel, J., & Daelemans, W. (1997). Memory-based learning: Using similarity for smoothing. In P. R. Cohen & W. Wahister (Eds.), *Proceedings of the 35th annual meeting of the Association for Computational Linguistics and eighth conference of the European chapter of the Association for Computational Linguistics* (pp. 436–443). Morristown, NJ: Association for Computational Linguistics.

Zhang, J. X., Feng, C., Fox, P. T., Gao, J., & Hai Tan, L. (2004). Is left inferior frontal gyrus a general mechanism for selection? *Neuroimage, 23*, 596–603.

Ziff, P. (1960). *Semantic analysis*. Ithaca, NY: Cornell University Press.

Figure captions

Figure 1. Effects of varying the lower limit neighborhood size on the two-alternative forced-choice decision accuracy for adjective ordering (Task 1), the ordering of adjectives with respect to other parts of speech in the noun phrase (Task 2), and the semantic similarity judgment task (Task 3 δ_{SM}). For comparison, the accuracy of the model using PMI to solve Task 3 is also plotted (Task 3 PMI). Note that this accuracy cannot be affected by neighborhood size as the formula for calculating the PMI value does not contain *k*. Note also that the x-axis uses a logarithmic scale.

Figure 2. Density plots (Sarkar, 2010) of the patient and control participant scores for Task 1. The plus signs and dashed line represent the patient scores (percentages of preferred order choices) and their density, respectively. The circles and full line represent the control group scores and their density. The cross represents the one patient (3297) who showed impaired performance on all three tasks and, hence, cannot be considered to be selectively impaired on Task 1. The vertical dotted line is Kemmerer et al.'s (2009) threshold for separating out the impaired patients (note that applying their criterion identifies nine impaired patients, as opposed to the seven patients they reported). The black and gray symbols represent the two clusters that the *k*-means clustering algorithm identified. The black plus signs form the group of selectively impaired participants in our analysis.

Figure 3. Predicted partial effect of breakdown *k* on selected adjective order and its interaction with participant group. The solid line represents the effect of breakdown *k* in the impaired group; the dashed line represents the breakdown *k* effect in the

unimpaired group. The gray areas are approximate 95% confidence bands. Log odds were transformed back to probabilities.

Table 1

Table 1. Mean accuracy scores (in percentage correct) of the unimpaired and the selectively impaired participants on the three 2AFC tasks of Kemmerer et al. (2009), and the corresponding model scores for the neighborhood sizes at which the model best approximates the participants' performance. The values for Task 3 within parentheses are the scores of the PMI approach. Note that the participant groups do not entirely coincide with those of Kemmerer et al. (2009), but are based on an automatic repartitioning of the participant scores (see Section 2.2.1).

Group		Task 1	Task 2	Task 3
		a big brown dog vs. *a brown big dog*	*a cool light rain* vs. *rain light cool a*	cue: *big*, target: *little*, distractor: *good*
Unimpaired	Participants	94.77	99.7	99.23
	Model (k = 1)	87.14	100	76 (90)
Impaired	Participants	69.71	97.32	95.2
	Model (k = 4,000)	67.14	100	72 (90)

Cells indicate mean accuracy (in %).

Table 2

Table 2. Predictors in the mixed logit model of selected adjective order, their measurement scale, and associated quantitative information.

Variable	Scale	Description
Fixed effects:		
Selected order	Dichotomous	*n*(preferred) = 2817, *n*(dispreferred) = 199
Participant group	Dichotomous	*n*(impaired) = 290, *n*(unimpaired) = 2726
Breakdown *k*	Ratio	*M* = 6327, *SD* = 10949.08, *min* = 2, *max* = 40000
Class distance	Ratio	*M* = 1.03, *SD* = 1, *min* = 0, *max* = 3
Rating difference	Ratio	*M* = 2.84, *SD* = 0.49, *min* = 1.3, *max* = 3.5
Random effects:		
Participant	Nominal	52 levels
Item	Nominal	58 levels

Of the 70 items in the data, naturalness ratings (from which the variable rating difference is derived) were available for 58 items. The regression analysis and the descriptive statistics in this table are therefore based on this subset.

Table 3. Estimated fixed effects of a mixed logit model predicting selected adjective order from breakdown *k*, rating difference, class distance, and participant group.

Parameter	*B*	*SE(B)*	*z*	*p*
Intercept	0.92	0.37	2.51	**.012**
Participant group	2.97	0.41	7.25	**< .001**
Rating difference	−0.19	0.33	−0.59	.554
Class distance	0.32	0.14	2.25	**.025**
Breakdown *k*	0.15	0.05	2.79	**.005**
Rating Difference × Participant Group	1.14	0.38	3.02	**.003**
Class Distance × Participant Group	0.63	0.21	3.00	**.003**
Breakdown *k* × Participant Group	−0.16	0.07	−2.10	**.036**

The *p*-values of significant effects are indicated in bold. All ratio variables in the model are mean-centered.

Table 4

Table 4. Nearest neighbor sets of the color adjectives for Task 1 of Kemmerer et al. (2009), at a neighborhood size of 5.

Adjective (frequency)	Nearest neighbors
black (4263)	black, white, red, brown, gray
blue (2309)	blue, red, yellow, gray, brown
brown (1811)	brown, gray, mother, horse, wife
gray (1158)	gray, brown, dark, blue, yellow
green (2910)	green, red, blue, seed, yellow
orange (520)	orange, grapefruit, fruit, lime, apple
purple (289)	purple, dress, chair, hat, peace
red (3752)	red, yellow, blue, white, green
white (5144)	white, black, red, yellow, gray
yellow (1433)	yellow, red, blue, brown, gray

The values between brackets are the adjective counts in the training corpus.

Figure 1

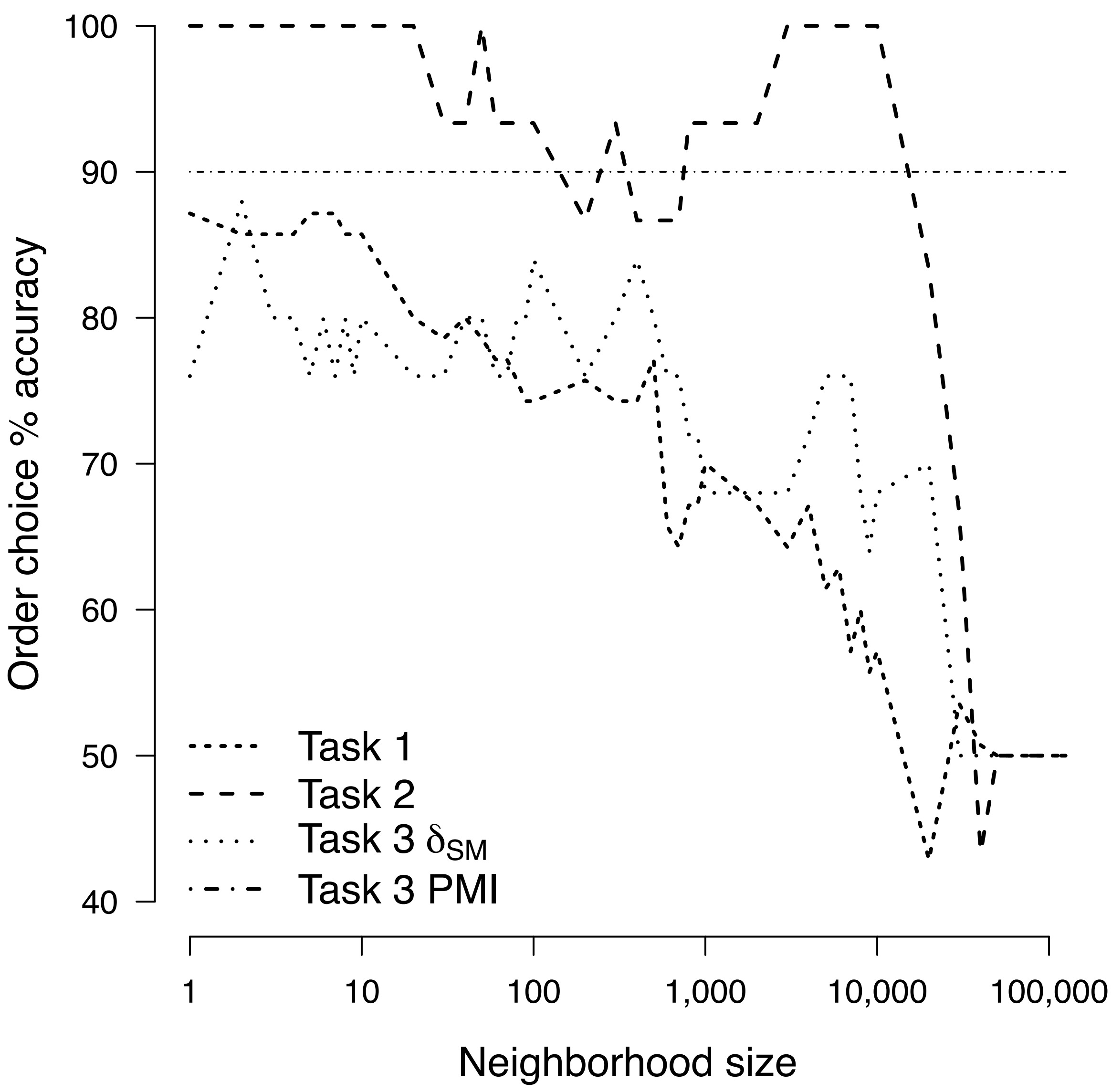

100
90
80
70
60
50
40
Order choice % accuracy
1
10
100
1,000
10,000
100,000
Neighborhood size
Task 1
Task 2
Task 3 δ_{SM}
Task 3 PMI

Figure 2

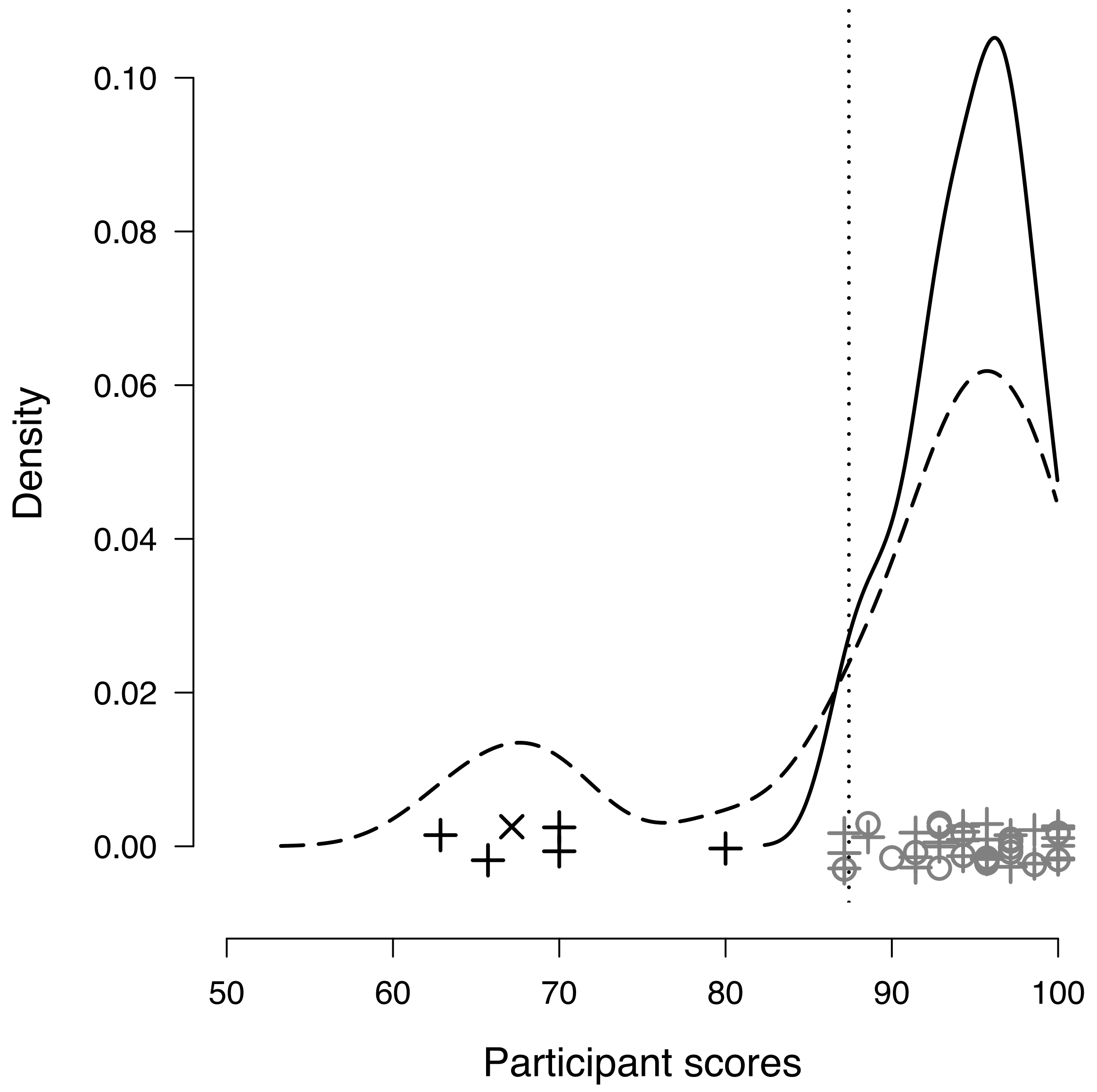
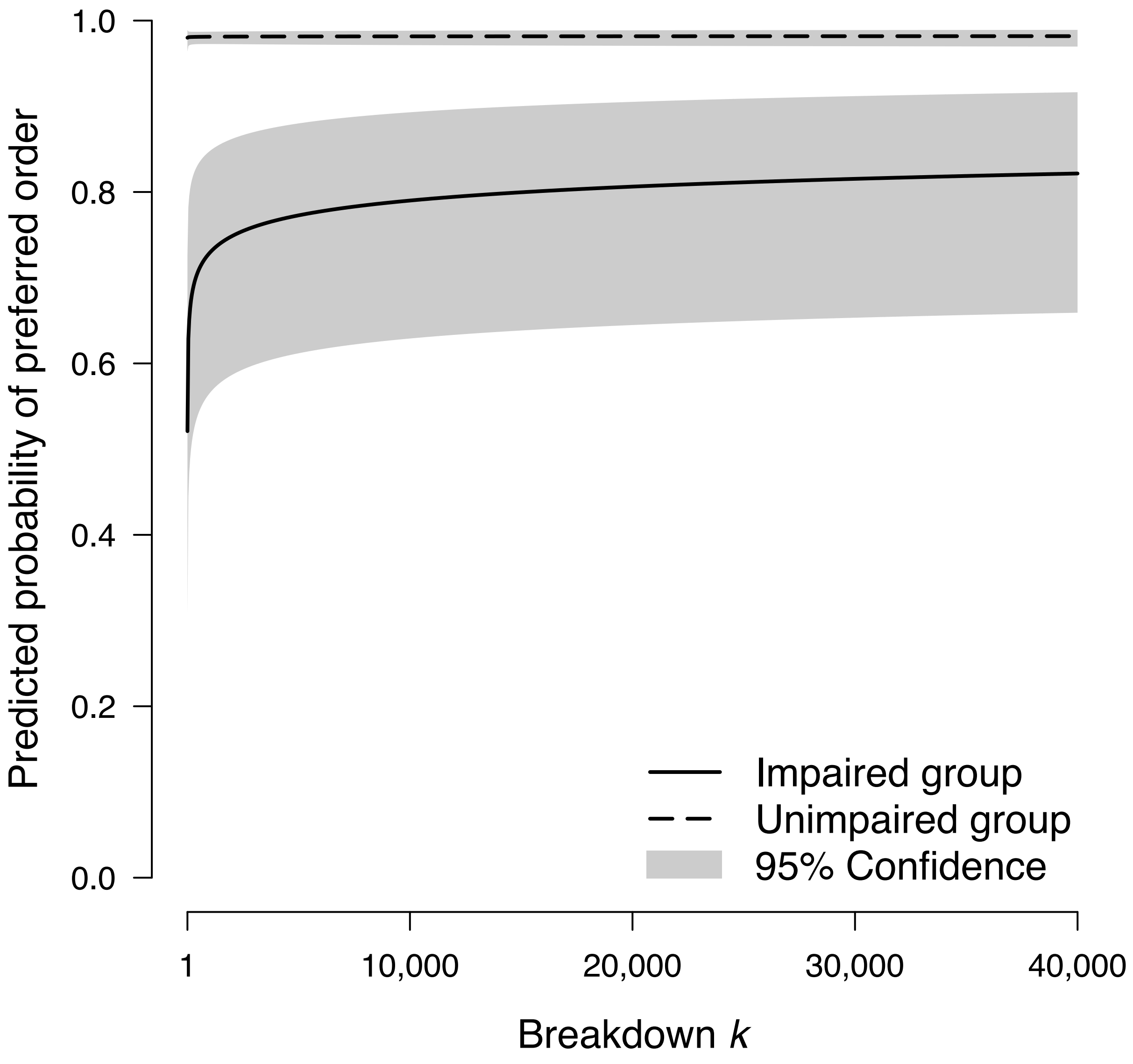

Density
0.10
0.08
0.06
0.04
0.02
0.00
50
60
70
80
90
100
Participant scores

Figure 3


Predicted probability of preferred order
1.0
0.8
0.6
0.4
0.2
0.0
1
10,000
20,000
30,000
40,000
Breakdown *k*
Impaired group
Unimpaired group
95% Confidence