A closer look at scalar diversity using contextualized semantic similarity[1]

Matthijs WESTERA — *Universitat Pompeu Fabra*
Gemma BOLEDA — *Universitat Pompeu Fabra / ICREA*

Abstract. We take a closer look at van Tiel et al.'s (2016) experimental results on diversity in scalar inference rates. In contrast to their finding that semantic similarity had no significant effect on scalar inference rates, we show that a sufficiently fine-grained notion of semantic similarity does have an effect: the more similar the two terms on a scale, the lower the scalar inference rate. Moreover, we show that a context-sensitive notion of semantic similarity (in particular ELMo; Peters et al., 2018) can explain more of the variance in the data, but only modestly, only for stimuli that contain informative context words, and only when the scalar terms themselves are sufficiently context-sensitive.

Keywords: scalar inference, scalar diversity, semantic similarity, relevance, distributional semantics, context.

1. Introduction

Scalar inference is the phenomenon whereby asserting a weaker proposition can warrant inferring the negation of certain stronger alternatives. To illustrate:

(1) a. It is warm. ⇝ It is not hot.
b. The boy dislikes broccoli. ⇝ The boy doesn't *loathe* broccoli.
c. The teacher believes it is true. ⇝ The teacher doesn't *know* that it is true.
d. The nurse saw some of the signs. ⇝ The nurse didn't see *all* of the signs.

When considering different words and constructions, the rate at which scalar inferences are drawn can vary greatly (Doran et al., 2009; van Tiel et al., 2016; Gotzner et al., 2018). Van Tiel et al. (2016) demonstrate such 'scalar diversity' experimentally with stimuli such as the following, asking participants for binary answers (*Yes/No*):

(2) John says: "The sand is warm". Would you conclude from this that, according to John, the sand is not hot?

They tested 43 weaker/stronger word pairs in this way, including those in (1), i.e., *warm/hot*, *dislike/loathe*, *believe/know*, and *some/all*. See figure 1 for all the word pairs they tested – most of them are adjectives, four pairs are closed-class items (*some/all*, *may/will*, *may/have to*, and *few/none*). Van Tiel et al.'s results across these various pairs comprised basically the full range between 0% and 100% of participants choosing *Yes*.

[1]This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 715154). This paper reflects the authors' view only, and the EU is not responsible for any use that may be made of the information it contains. This project has also received funding from the Ramón y Cajal programme (grant RYC-2015-18907) and from the Catalan government (SGR 2017 1575).

 In: M. Franke et al. (eds.)
Proceedings of Sinn und Bedeutung 24, vol. 2, pp.439–454. Osnabrück University.

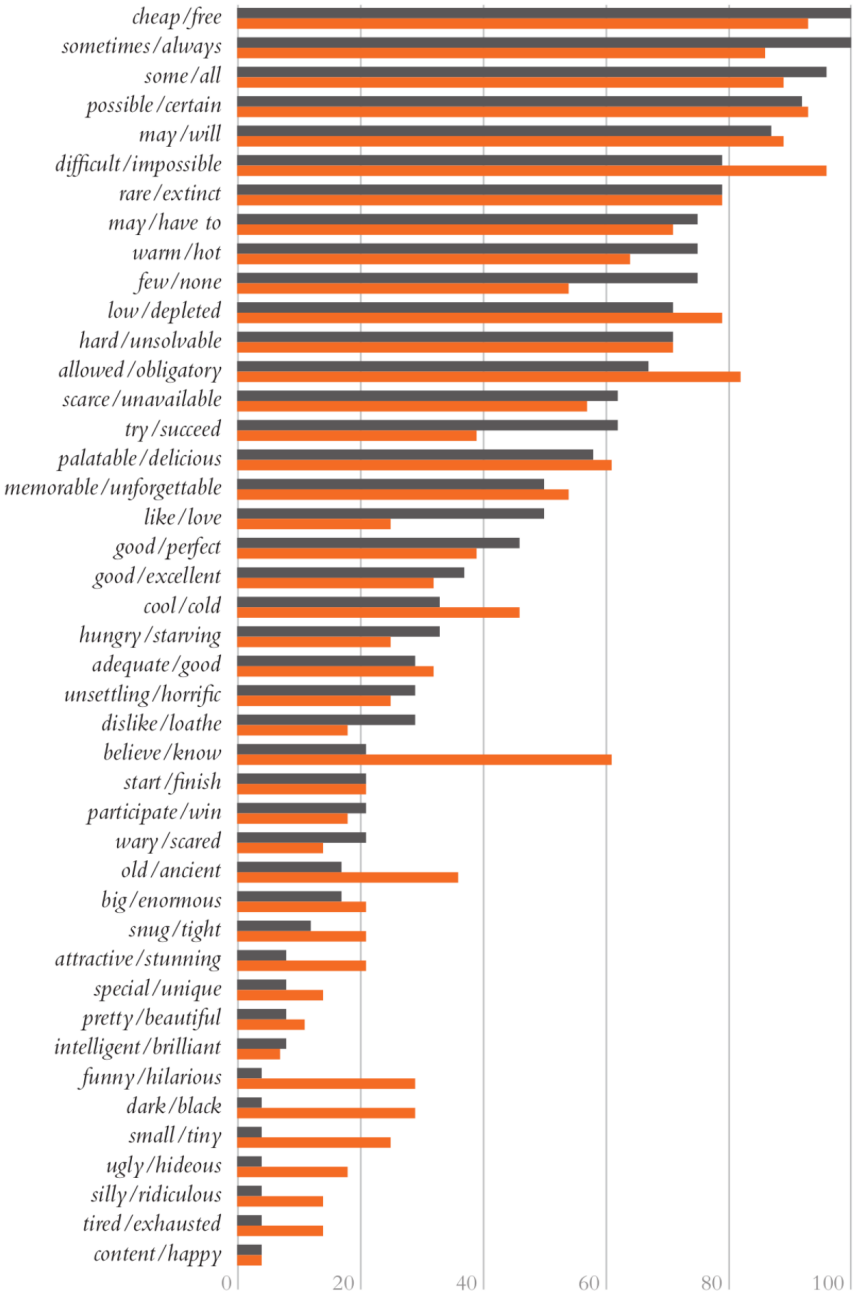

Figure 1: Figure taken from van Tiel et al. 2016: the percentage of participants responding "yes" for each item, in Experiment 1 (gray) and Experiment 2 (orange).

The aim of this paper is to contribute to explaining the scalar diversity in van Tiel et al.'s data, by taking a closer look at the role of *semantic similarity* in the results. Van Tiel et al. note that semantic similarity ought to play a role in explaining scalar diversity, but they find no effect. In line with observations by McNally (2017), we suspect that a more fine-grained notion of semantic similarity than the one considered by van Tiel et al. might do a better job. More precisely, the notion of semantic similarity used by van Tiel et al. (Latent Semantic Analysis; LSA; Landauer and Dumais, 1997a) is context-independent: it assigns the same number to a pair of words (representing their similarity) regardless of the context in which they occur. Because as McNally notes context may matter in various ways, we investigate whether perhaps a context-sensitive notion of semantic similarity, obtained from the more recent ELMo model (Embeddings from Language Models; Peters et al., 2018), could do a better job at explaining scalar diversity.

2. Background

2.1. Scalar inference

Different accounts exist of what causes scalar inferences to arise. A common explanation is pursued in the so-called *neo-Gricean approach* (see Geurts, 2011 for a critical but ultimately favorable discussion): if the speaker in each of the examples in (1) had believed the stronger proposition, they would have asserted that instead – since they did not, they must believe it is false. Another is pursued in the so-called *grammatical approach*: the weaker and the stronger proposition stand in a certain grammatical relationship that guarantees that the statements in (1) (on the left of the $\rightsquigarrow$) are ambiguous between a reading with and a reading without the scalar inference, and when facing ambiguity we would simply choose the strongest interpretation. A more recent proposal is *attentional pragmatics* (Westera, 2017), which maintains that scalar inference arises not from stronger alternatives not being asserted, but from them not even being *mentioned* (see also Westera, 2020 in this volume).

The discussion in van Tiel et al. (2016) are grounded primarily in the neo-Gricean approach, but not in a way that restricts their conclusions (or ours in the present paper) to that branch. All approaches to scalar inference are compatible in principle with scalar diversity as observed in van Tiel et al. (2016), though some more explicitly so than others. For one, each approach in principle permits the existence of many other pathways to scalar inference (Geurts, 2011), e.g., lexical semantics, typicality inference, and various pragmatic routes. Moreover, even within the pathways favored by each approach there exist parameters representing contextual relevance, lexical knowledge and general world knowledge, each of which can influence the degree to which a scalar inference is predicted for a given example. Though while each theory permits scalar diversity, explaining it is another matter: contextual relevance, lexical knowledge and general world knowledge are each notoriously difficult to model in their own right.

2.2. Van Tiel et al. (2016): Scalar diversity

Van Tiel et al. (2016) show experimentally that the perceived presence of *scalar inferences* varies greatly between different scales (see also Doran et al.; Gotzner et al., 2009; 2018), with stimuli such as the following for the scale *warm*/*hot*:

(3) John says: "That is warm". Would you conclude from this that, according to John, it is not hot? Yes/No.

They perform two experiments with 25/30 participants each, each with the same 43 pairs of words, such as *warm*/*hot*, *adequate*/*good* and *believe*/*know*. In Experiment 1 the context words in the stimuli are minimally informative, containing pronouns such as "that" in (4), whereas Experiment 2 contains more descriptive context words, e.g., "this sand" in (4):

(4) John says: "The sand is warm". Would you conclude from this that, according to John, the sand is not hot? Yes/No.

The context words van Tiel et al. used in Experiment 2 were obtained experimentally by asking 10 participants to fill in the blanks in sentences such as the following:

(5) The _____ is warm but it isn't hot.

From the resulting 10 candidate expressions per item, van Tiel et al. selected 3 expressions for each item based on two constraints: try to select two frequent and one infrequent choice, and try to ensure some diversity in the range of expressions. The three selected expressions for each item were used as context words in the stimuli in Experiment 2. For the item *warm*/*hot*:

(6) a. The weather is warm. b. The sand is warm. c. The soup is warm.

Van Tiel et al.'s results in both Experiment 1 and 2 comprised essentially the full range between none and all of the participants choosing *yes*. This is shown in figure 1. Van Tiel et al. report that, overall, the rates of "yes" responses did not differ significantly between the two experiments, and that there was no pair of stimuli for any scale that differed significantly in this regard, either. Van Tiel et al. consider two broad factors for explaining the variation in scalar inference they find: *availability* of the stronger item on the scale as a relevant alternative, and *distinctness* of the two items on the scale.

The **availability** of the stronger term as a relevant alternative conceivably affects scalar inference because, according to most theories, scalar inference is the exclusion of relevant alternatives. If the stronger term on a scale is not in fact a relative alternative, then the reason why the speaker did not use it is that it is irrelevant, not that it is false. This is conceivably the case for the scale *participate*/*win* (though this is not the explanation we think the data ultimately favors): upon hearing (7) one may not normally guess that the question of whether she won was likewise relevant – perhaps the question of participation and the question of winning are normally considered one at a time:

(7) She participated. ($\not\rightsquigarrow$ she didn't win)

The unavailability of "she won" as a relevant alternative could explain the low rate of scalar inference for this scale. More generally, the expectation is that the more readily available the stronger term is as a relevant alternative, the higher is the rate of scalar inference.

The other factor, **distinctness**, conceivably affects scalar inference because if the two terms on the scale are insufficiently distinct, then the speaker's choice for one rather than the other may well be arbitrary (for present purposes) or due to imprecision. In that case one cannot conclude from the speaker's use of the weaker term the negation of the stronger term. Perhaps

this is illustrated by the scale *special/unique*, since the types of contexts where the difference between these would matter (i.e., contexts where one could reasonably say "It is special, but not unique") seem to us quite atypical:

(8) It is special. ($\not\rightsquigarrow$ it isn't unique)

Thus, if the weaker and the stronger term are not relevantly distinct, one expects a lower rate of scalar inference. Availability and distinctness are in a way opposite forces: the two terms on a scale should be related (lest the stronger one will not be available as a relevant alternative), but not too similar (lest the difference between them be irrelevant or negligible).

Van Tiel et al. consider a number of variables which they suspect may correlate with availability and distinctness. We refer to van Tiel et al.'s own discussion of these factors as well as the discussion in McNally (2017), and concentrate here on only one factor: **semantic similarity**, which they take from the distributional semantic approach Latent Semantic Analysis (LSA; Landauer and Dumais, 1997b; see below). Van Tiel et al. consider semantic similarity primarily as a measure of availability: for the stronger term to be readily available as a relevant alternative to the weaker term, the two terms must tend to be relevant in similar contexts, which means they tend to be used in similar contexts – and words with similar contexts of use are assigned similar semantic representations in distributional semantics (see section 2.4). However, we think semantic similarity should also correlate (inversely) with distinctness: the more semantically similar the two terms on a scale, the less distinct they are.

The expectation for scalar inference in general is that words must be similar (lest the stronger term not be available as a relevant alternative) but not too similar (lest they be insufficiently distinct for the difference between them to matter). Contrary to expectation, van Tiel et al. find no evidence for this hypothesis; they find no effect of semantic similarity. It seems unlikely that the positive and negative effects of semantic similarity would exactly cancel each other out (and there are reasons to believe only one of the two factors is active anyway; see section 6). So why do they not find any effect? The aim of the present paper is to understand this better, and in particular to see whether this reflects a shortcoming of the notion of semantic similarity used and whether a better notion exists for which the data do show an effect.

2.3. McNally (2017): Context matters

Our aim to understand the absence of an effect of semantic similarity in van Tiel et al.'s data is shared by McNally (2017). McNally notes that the notion of semantic similarity used by van Tiel et al. (based on a particular distributional semantic model, see section 2.4) is rather coarse-grained. More precisely, it is context-independent, in the sense that the same vector is assigned to a word regardless of the sentence in which it occurs. Accordingly, it may not do justice to the particular senses with which the terms are used in their experiment. As McNally notes (p.5), "though *warm* and *hot* are scalemates for ascribing temperature, *hot*, but not *warm*, is used for popularity (*a hot/??warm product*), temper, and sex appeal (*a hot/??warm body*); while *warm*, but not *hot*, is used for friendliness or empathy (*a warm/??hot personality*)." As a consequence, the overall semantic similarity of the two terms *warm* and *hot* will underestimate their actual similarity in the temperature ascriptions that comprise van Tiel et al.'s stimuli.

Besides context narrowing down the sense in which both terms of a scale are used, thereby rendering the scalar inference rate *higher* than one would expect on the basis of a context-independent notion of semantic similarity, context can in principle also favor interpreting the two terms of a scale in two different senses, leading to a *lower* scalar inference rate. For instance, McNally suggests that (9) may have a low scalar inference rate because the adequacy/goodness of one's salary can be assessed in different ways:

(9) The salary is adequate. ($\not\rightsquigarrow$ the salary is not good)

Participants in van Tiel et al.'s experiment were free to interpret "adequate" in one way (e.g., as meeting one's needs) and "good" in another (e.g., being better off than one's peers), which strike us as reasonable interpretations. But because the goodness of one's salary compared to peers is a separate question from its adequacy in meeting one's needs, under this interpretation "good" (in the sense of compared to peers) is not available as a relevant alternative to "adequate" (in the sense of meeting one's needs). Accordingly, participants who favored this interpretation would report no scalar inference.

Context can also affect the availability of a relevant alternative by modulating expectations of relevance itself (i.e., not just by favoring particular senses for the two terms, as in (9)). McNally suggests that (10a) may imply "not hot" because hot sand can be dangerous and hence relevant, whereas (10b) may lack this implication because it may well be used instead to contrast warm (and hot) soups with *cold* soups instead:

(10) a. The sand is warm. ($\rightsquigarrow$ the sand is not hot)
b. The soup is warm. ($\not\rightsquigarrow$ the soup is not hot)

These and the foregoing examples are mere illustrations of how context could in principle influence scalar inference. As it turns out, in van Tiel et al.'s data, example (9) indeed seems to behave in the way suggested by McNally, but (10) does not – but these particular examples do not really matter for present purposes: the main point, that there are ways for context to influence scalar inference, holds regardless. We agree with McNally (and, e.g., Geurts (2011), whom she notes holds a similar view) that it is important in general not to loose track of the ways in which context, lexical knowledge and world knowledge can affect phenomena such as scalar inference. The more immediate takeaway for present purposes is that a context-independent notion of semantic similarity, such as that used by van Tiel et al., may be too coarse-grained to adequately model effects of semantic similarity (availability, distinctness) on scalar inference.

2.4. Semantic similarity

The notion of semantic similarity considered by van Tiel et al. (2016), as well as the additional notions we will consider in this paper, are derived from *distributional semantics*. Distributional semantics is based on the 'distributional hypothesis' (Harris, 1954), which states that words with similar meanings are used in similar kinds of contexts (i.e., have similar distributions). In distributional semantics, the meaning of a word is represented as a high-dimensional numerical vector, derived by abstracting over occurrences of the word in large amounts of data.

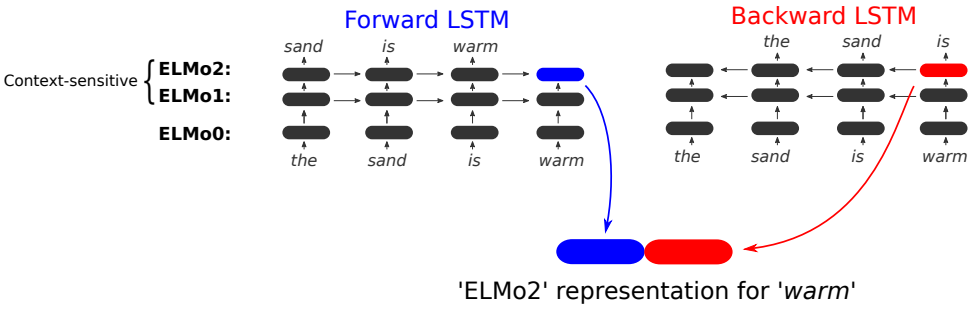

Figure 2: Schematic representation of the ELMo model (Embeddings from Language Models) applied to the sentence "the sand is warm".

These vector representations have been shown to correlate with many aspects of word meaning, and they have been successfully used as models of semantic similarity in cognitive science and computational linguistics (for overviews see Clark (2015); Lenci (2018); Boleda (2020)). Semantic similarity between two words is computed as the cosine of the angle between their vectors: the cosine is 0 when the vectors are exactly orthogonal, 1 when they point in the same direction, and -1 when they point in opposite directions.

Traditional 'count-based' methods of distributional semantics start from a (huge) table of word-context occurrence counts and apply dimensionality reduction to derive the word vectors. More recent 'prediction-based' methods instead start from random word vectors (typically the weights in an artificial neural network) and incrementally update them to better predict word-context occurrences (for a comparison see Baroni et al., 2014). Van Tiel et al. (2016) obtained their notion of semantic similarity from an influential count-based model: Latent Semantic Analysis (LSA; Landauer and Dumais, 1997b), in particular the model referred to as "General Reading up to 1st year college" available at `http://lsa.colorado.edu`. In this paper we will compare this notion of semantic similarity to notions obtained from three alternative distributional semantic models: the count-based model GloVe (Global Vectors; Pennington et al., 2014), and the prediction-based models Word2Vec (Mikolov et al., 2013) and ELMo (Embeddings from Language Models; Peters et al., 2018).

Standard distributional semantic models (both count-based and prediction-based), including LSA, Word2Vec and GloVe, assign a single vector to each word in the vocabulary. These models do not come equipped with a systematic method for 'contextualizing' these representations, i.e., for assigning (slightly) different vectors to the same word in different contexts, vectors which could represent the particular 'sense' in which the word is used in that context. ELMo's main innovation, and the engine behind its enormous success on many NLP tasks, is that it does come with such a method. Because the resulting context-dependence of its vector representations is crucial for present purposes we will briefly summarize how this works. We will not summarize the other models: we include LSA merely as a sanity check (i.e., comparison to van Tiel et al.'s results), GLoVe merely as a more recent representative of the count-based models of which LSA is also a specimen, and Word2vec because it was effectively the state-of-the-art in neural network models of distributional semantics prior to ELMo.

ELMo is a neural network model, schematically depicted in figure 2. It takes as input a sentence (in the figure "the sand is warm"), one word at a time, and is trained on the task of predicting

the next word at every step. The part relevant for present purposes can be broken up into two sub-models which traverse the sentence in opposite directions (to account for the fact that the contextualized meaning of a word can depend both on what comes before and on what comes after). Each sub-model consists of three layers. The first layer (in fact shared between both sub-models) assigns a context-independent vector representation to each word in the vocabulary, comparable to those of the other models of distributional semantics. In the next two layers these representations are iteratively combined with a vector representation of the context (from the left or from the right), using a recurrent neural network of the influential Long Short-Term Memory type (LSTM; Hochreiter and Schmidhuber, 1997). The contextualized representation of a word computed by ELMo at a given layer is the concatenation of its representations in the forward and backward sub-models at that layer. Normally, the contextualized representations used for downstream tasks are computed as weighted averages of the three layers (but see below).

3. Approach

In the previous section we reviewed why semantic similarity is expected to have an effect – positive (availability) or negative (distinctness) – but also why an appropriate notion of semantic similarity should be context-sensitive. To test whether this expectation is borne out we fit a number of logistic regression models on the data from van Tiel et al. (2016) (the individual responses). Each model is fitted to predict the yes/no responses as the dependent variable, based on a notion of semantic similarity as independent variable, where different models use different notions of semantic similarity: some context-independent and some context-dependent. We compare models by considering both the percentage of variance they explain (pseudo-R^2) and the effect size (β).[2] We do this for the data from van Tiel et al.'s Experiments 1 and 2 separately, in order to see whether the different degrees of contextualization in their respective stimuli has an effect.

We consider six notions of semantic similarity. Each is computed as the cosine similarity between vector representations of words, where the vector representations come from different distributional semantic models as described in the previous section:

- **LSA:** the classical, count-based distributional semantic model used by van Tiel et al., for which they found no effect;
- **GloVe:** A more recent, count-based distributional semantic model, whose representations are likewise context-independent;
- **Word2vec:** An influential neural-network based distributional semantic model, whose representations however are still context-independent;
- **ELMo0:** The first word embedding layer of ELMo, which is context-independent;
- **ELMo1:** The second layer of ELMo (the first layers of its recurrent modules, concatenated), which is context-dependent;
- **ELMo2:** The third layer of ELMo (the second layers of its recurrent modules, concatenated), which is likewise context-dependent.

[2] We fit logit models using Python's `statsmodels` package (Seabold and Perktold, 2010).

	LSA	Word2vec	GloVe	ELMo0	ELMo1	ELMo2
That is warm/hot:	0.51	0.432	0.745	0.622	0.805	0.844
The weather is warm/hot:					0.807	0.839
The sand is warm/hot:		—— same ——			0.815	0.847
The soup is warm/hot:					0.811	0.856

Table 1: Similarity scores for the scale *warm/hot* and its various stimuli. Beware that the similarity scores assigned by different models cannot be directly compared.

For LSA, GloVe, Word2vec and ELMo0, obtaining the word representations for the terms in each scale (on which then to compute cosine similarity) is a matter of looking them up in each model's list of word representations. For ELMo1 and ELMo2 the model is instead given the whole stimulus sentence (e.g., "the sand is warm") after which the representation of the scalar term "warm" in the two layers of the recurrent module is extracted. Thus, whereas from the context-independent models we obtain one number per scale representing the semantic similarity of the scalemates (e.g., one number for *warm/hot*), and likewise for the first layer of ELMo, from the context-dependent ELMo layers ELMo1 and ELMo2 we obtain one number per sentential stimulus, i.e., four numbers per scale (one from Experiment 1 and three from Experiment 2). For instance, for the scale *warm/hot* we get one number for *it is warm/it is hot* from Experiment 1, another for *the sand is warm/the sand is hot* from Experiment 2, and two more numbers for variants of the latter. See table 3 for concreteness.

Strictly speaking, extracting the individual layers of ELMo and using them separately is not how ELMo is normally used, or how it was intended. Normally one would use a weighted sum of the three layers, where the weights are finetuned to a particular task. The idea behind this is that different layers will likely encode different kinds of information about the word – and indeed, Aina et al. (2019) show that whereas the ELMo1 representations are still strongly based in the current word, the ELMo2 representations contain more information about the *next* word (i.e., the word to be predicted by the model during training) – which makes sense given the task on which ELMo is trained (see again figure 2). We did fit a model of this kind (i.e., with all three layers as independent variables) but found no real improvement in R^2 over the models using just ELMo1 (moreover, because the number of scales used in van Tiel et al. (2016) is quite small (43) we fear the added parameters of this approach increase the risk of overfitting). Accordingly, below we will report only the results of the models based on individual ELMo layers. Note that such a within-ELMo comparison is interesting for the current research question in its own right, given the different degrees of 'contextualization' in the different layers (Aina et al., 2019).

4. Results

Figure 3 shows the percentage of the variance explained (pseudo-R^2) by each of the fitted models. It reveals that the context-independent models LSA, word2vec and Glove capture hardly any of the variance in the data, in line with the findings of van Tiel et al. (2016). The ELMo models fit the data better, explaining up to 6% of the variance in Experiment 1 and up to 4% in Experiment 2. Among the ELMo models, on Experiment 1 the context-independent model ELMo0 is best, followed by the context-dependent models ELMo1 and ELMo2; on Experiment 2, context-dependent ELMo1 takes the lead.

Figure 3: Percentage of variance explained (Pseudo-R^2) by the fitted models.

Figure 4: Coefficients of the fitted models.

Figure 4 shows the coefficients of the fitted models. The similarity scores were normalized prior to model fitting (divided by standard deviation), so the exponentials' coefficients are interpretable as the expected change in the odds of a "yes" response (i.e., a scalar inference) if the similarity is increased by one standard deviation. The plot shows bigger (negative) effects for the ELMo models, with, e.g., in Experiment 2, increasing the ELMo1-similarity by one standard deviation yields around a 40% ($1 - e^{-.5}$) decrease in the odds of a "yes" response. For LSA the effect seems considerable as well, though in the opposite direction – but recall that it explains much less of the data. Word2vec and GloVe have hardly any effect, suggesting that the worse performance of the LSA model compared to the ELMo models is not due to LSA simply being an older model or due to it not being a neural network model. Among the ELMo models, the coefficients show the same ranking as the R^2 in figure 3: on Experiment 1 the context-independent layer ELMo0 has the largest effect, followed by ELMo1 and ELMo2; on Experiment 2, context-dependent ELMo1 takes the lead.

We noticed that ELMo assigns considerably lower similarities to the four closed-class items

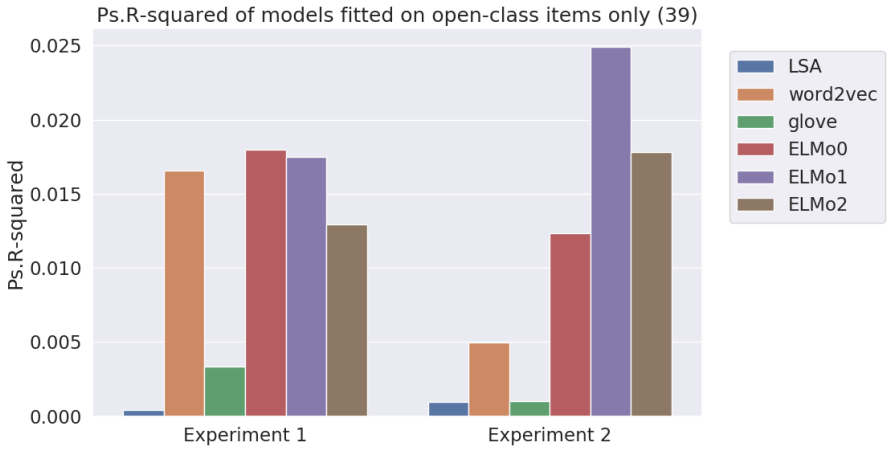

Figure 5: Percentage of variance explained (Pseudo-R^2) by the fitted models, restricted to the 39 open-class items.

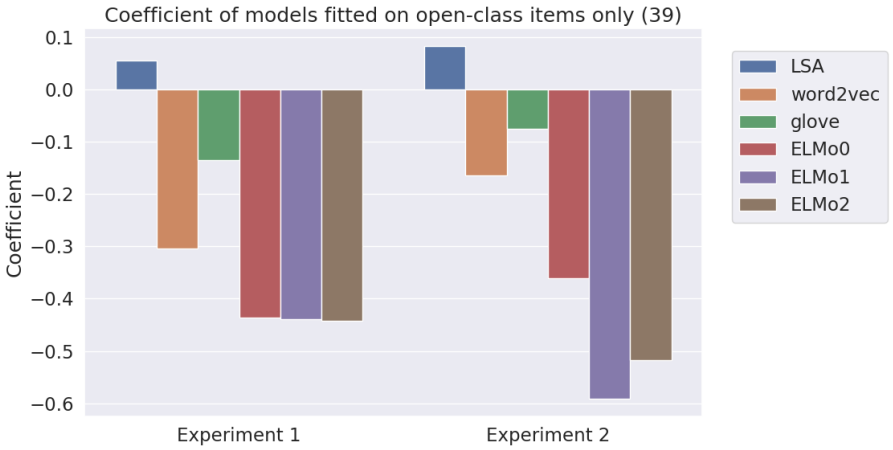

Figure 6: Coefficients of the fitted models, restricted to the 39 open-class items.

compared to the open-class items (mostly adjectives): *some/all*, *may/will*, *may/have to*, and *few/none*. Since these items are also more likely to trigger a scalar inference, that could explain why semantic similarity in the ELMo models has a negative effect – perhaps these items dominate the data. To find out whether the effect persists also within the (more uniform) subset of open-class items, we fitted the same models on just those (39 out of 43 items). Figures 5 and 6 show the pseudo-R^2 and coefficients of the resulting models. Figure 5 reveals that a substantial portion of the variance explained by the original ELMo models for all items (compared to Figure 3) must have been due to the closed-class items: on the open-class items pseudo-R^2 drops to a mere 1.5%. Another striking difference compared to the earlier figures is that, on the open-class items, Word2vec is somehow doing a lot better in Experiment 1 (though still not in Experiment 2). Besides these differences the same crucial pattern is visible among the ELMo models: in Experiment 1 context-dependence appears not to make a difference (ELMo0 and ELMo1 perform alike) while in Experiment 2 it does, with ELMo1 being the winner, even by a slightly larger margin than before, reflecting perhaps that the open-class items are more

context-dependent than the closed-class items.

5. Discussion

Let us explore some possible explanations for the foregoing observations, some more speculative than others, and all in need of further investigation.

The ranking of the models (in terms of variance explained, i.e., figures 3 and 5) suggests that the ELMo models provide the best word representations for predicting an effect of semantic similarity on scalar inference, *regardless* of the hypothesized effect of context: the context-independent ELMo0 model performs better (Experiment 1) or almost as well (Experiment 2), after all. We initially thought that the general superiority of ELMo could reflect its better grasp on closed-class items (function words), and this is certainly true compared to Word2Vec, which improves to almost the level of ELMo when considering only open-class items (figure 5, left). Indeed, closed-class items are known to be challenging for (especially traditional) distributional semantic models, which have tended to focus on open-class items from the outset both in training and in evaluation methods. But LSA and GloVe do not improve when considering only the open-class items, suggesting that there is something else at play, too.

A potentially relevant difference between the four models in this regard is the *context window* used during training, on which they form a kind of spectrum: for LSA the context window during training is a full document, for ELMo it is a (potentially long) sentence, for Word2Vec it is a set of neighboring words, and for GloVe it is a single neighboring word. It is conceivable that the large context window of LSA makes it too coarse-grained for subtle meaning distinctions such as between the two terms on a scale: e.g., from the fact that *warm* and *hot* occur in similar documents (say, documents involving cooking, the weather, clothes) the LSA model may learn that both scalar terms relate to things having a temperature, but perhaps not their precise distance on the temperature scale. For GloVe we conjecture that it is, rather, its *small* context window that is the problem: it conceivably makes GloVe less able to detect meaning differences that affect not the direct neighbors of words, but only their indirect (longer-distance) neighbors – and the scalar differences in the current data may be of that kind. To illustrate, the scalar terms *hot* and *warm* may (in their temperature sense) apply to exactly the same kinds of things (soups, weather, clothes, etc.), so perhaps the subtle scalar difference between them tends to affect only more indirectly related events (such as decisions of whether to start eating, whether to go for a hike, and what to wear), events which may tend to be described in the same (part of a) sentence as the scalar term (hence detectable by ELMo and Word2Vec) but typically not within one word distance of it (hence undetectable by GloVe). This would explain why GLoVe does not improve when considering only the open-class words, despite these being the types of words that distributional semantics is traditionally supposed to be good at. The foregoing is very speculative, and there are many important differences between the various distributional semantic models that we have not mentioned, but the context window size during training is the only difference that currently strikes us as at least potentially relevant to the issue at hand.

We may tentatively extend the foregoing line of explanation in terms of context window size to the (presumed) difficulty of closed-class items for LSA, Word2Vec and GloVe (i.e., their low R^2 in figure 3). Concerning LSA, since closed-class items are frequent in any document, on any topic, LSA's context of a full document makes the model too coarse-grained to tell such words apart. As for Word2Vec and GloVe, their context windows may rather be too small: it

is only in the context of an argument, i.e., at the level of sentences – the context window with which ELMo is trained – that the choice of one closed-class item over the other on the same scale (e.g., "may" instead of "will") tends to have an effect. Summing up, the preceding two paragraphs tentatively explain why only ELMo has a grasp of closed-class items, and why even on the open-class items only ELMo and Word2Vec can handle the subtle scalar differences involved in van Tiel et al.'s data, with LSA and GloVe struggling in both respects.

Next, the ranking among the ELMo-based models (in both figure 3 and figure 5) suggests that context, to the extent that it matters at all, offers at best only a marginal improvement (ELMo1 compared to ELMo0), and only in Experiment 2. The lack of an effect in Experiment 1 may not be too surprising, given that context was not particularly informative in Experiment 1 to begin with; only Experiment 2 featured informative context words, after all. Nevertheless, it is not entirely *un*surprising either: the stimuli of Experiment 1 do provide context in the form of syntactic structure, e.g., the stimuli reveal that a scalar term is used (say) predicatively as opposed to adjectivally, which should conceivably have an effect on scalar inference (e.g., because a predicate is more likely to be the information structural focus of the sentence).[3] Assuming that van Tiel et al.'s participants did pick this up, the fact that the context-sensitive layers of ELMo do not perform better than ELMo1 in Experiment 1 suggests that the ELMo model's word representations are not (sufficiently) affected by syntactic (plausibly information structural) context. Assuming that the problem is not ELMo itself (given its success on many NLP tasks), a plausible explanation for the latter could be that most of the scalar terms in the experiment are predominantly used predicatively anyway, such that this use would already dominate the (non-contextualized) word vector to begin with. We leave testing this conjecture to future work.

As for Experiment 2, where the stimuli contained more informative context words, what may be surprising there is how small a difference context seems to make, i.e., the small magnitude of the advantage of ELMo1 over ELMo0 (still figure 3). Again, let us assume that ELMo (given the success of ELMo on many NLP tasks) is in principle able to properly model the effect of context on the interpretation of the scalar terms. What the small magnitude of its effect then suggests is that only some of the scalar terms in the experiment were significantly context-sensitive to begin with. This is corroborated by the fact that figure 5, which considers only the open-class items, shows the same (or in fact slightly larger) absolute advantage in Experiment 2 for ELMo1 over ELMo0: since open-class items are generally more context-sensitive than closed-class items, the fact that ELMo1's advantage over ELMo0 resides mostly there suggests that the context-sensitive layers have an advantage only for context-sensitive words – and perhaps even among the open-class items there were not enough of those for context to have a bigger effect. Related to the effect of context, note that although Word2Vec performed quite well when restricted to the open-class items (figure 5), no such boost was observed in Experiment 2. Together with the relative performance of ELMo0 vs. ELMo1 in Experiments 1 and 2, this suggests that for Experiment 1 it may be sufficient for the model to have adequate context-independent representations of the scalar terms, but that for Experiment 2 context-sensitivity is required.

Next, what might we conclude from the fact that the best models (ELMo) show a *negative* effect of semantic similarity (figures 4 and 6)? Recall from section 2 that van Tiel et al. (2016)

[3]Thanks to Richard Breheny for bringing this possible factor to our awareness.

identified two possible influences on scalar inference, namely *availability* of the stronger scalar term as a relevant alternative (a positive effect), and the *distinctness* of the two scalar terms (a negative effect). Accordingly, the negative effect of our best models suggests that distinctness has a role to play in the experiment, but not availability, or at least less so. A possible explanation for this is the following. Although in general one would expect both availability and distinctness to affect the rate of scalar inference, in the scope of van Tiel et al.'s experiments perhaps only distinctness has a role to play. This is because the availability of the stronger terms as relevant alternatives may have been sufficiently fixed already by the experimental setting itself, a possibility considered also by van Tiel et al.: the experiment itself would imply the stronger term's availability as a relevant alternative, by virtue of explicitly asking participants about the scalar inference (see (3)). See Schwarz (1996) for discussion of this type of influence of experimental context on pragmatic assumptions.

Van Tiel's experiments also contain some other tentative evidence for the foregoing explanation. Recall that the stimuli in Experiment 2 were constructed by eliciting context words from participants that fit in the type of scheme in (5), i.e., words that make a sentence affirming the weaker term and denying the stronger term a natural thing to say. If availability had played a role, one would expect that using these words instead of the uninformative context words (e.g., pronouns) in Experiment 1 would overall increase availability and thereby scalar inference. But this is not what van Tiel et al. find: overall, scalar inference rates were not significantly higher in Experiment 2 than in Experiment 1. Given the diversity of van Tiel et al.'s stimuli, it seems unlikely that the (uninformative) stimuli of Experiment 1 would all independently already happen to favor, as their most typical interpretation, one which made the scalar alternative available; it seems more plausible that this was enforced by the experimental setting itself. Moreover, van Tiel et al. report that within Experiment 2 there were no significant differences between the stimuli for a given scale, e.g., "the sand is warm" and "the soup is warm" have the same scalar inference rate (contrary to the example of McNally, 2017), even though van Tiel et al. tried to opt for some diversity in context words, pointing again towards the absence of a positive effect of context words on availability across the board.

Zooming out a little, the percentage of variance explained by the various models seems to us rather small, especially for the models fitted on the open-class items only. Since our research question is primarily about the comparison of context-independent and context-dependent notions of semantic similarity, the absolute performance of the models does not immediately matter for the purposes of this paper. However, it does raise the issue of what the current models are missing. There seem to be three main possibilities:

1. Other factors, besides distinctness (and maybe availability), affect scalar inference in van Tiel et al.'s experiments;
2. Other factors besides contextualized semantic similarity affect distinctness (and maybe availability) in van Tiel et al.'s experiments;
3. The notion of contextualized semantic similarity as we extracted it from the ELMo model is not good enough for present purposes.

Exploring items 1 and 2 would be a substantial inquiry in its own right, which we leave to future work (see the discussions in van Tiel et al., 2016 and McNally, 2017 for some suggestions).

Item 3 bears more directly on the aims of this paper, and we will briefly explore it.

Item 3 is a live possibility not just because models are never perfect (and in future work we hope to test the contextualized word representations of models that have more recently beaten ELMo on other tasks), but also because the particular way in which we employed ELMo may not be the most suitable. To understand the latter, let us distinguish three aspects of context: (i) the words used in the stimuli's sentences besides the scalar term, (ii) the syntactic structure in which the scalar term appears (e.g., predicative vs. adjectival), and (iii) the experimental setting, where each stimulus explicitly asks whether the negation of the stronger scalar alternative can be inferred. In the way in which we applied ELMo to the stimuli, it can be sensitive to (i) and (ii), but not to (iii). To illustrate, recall that we used ELMo to compute the semantic similarity of "warm" and "hot" in the context of a sentence such as "the soup is ...", which does not tell ELMo that these are stimuli in an experiment that explicitly relates "warm" and "hot". To handle (iii) better, one could try to take the ELMo representations from sentences containing both scalar terms, such as "the soup is warm but not hot", which, although deviant from the stimuli used by van Tiel et al., at least tells ELMo that the context is one where both scalar terms occur. We leave this to future work.

6. Conclusion

Scalar inference – the inferred negation of a stronger statement from the utterance of a weaker statement – is expected to depend on the semantic similarity between the stronger term and the weaker term: the two terms should be similar (lest the stronger term not be available as a relevant alternative) but not too similar (lest a speaker's choice for one over the other be due to, e.g., imprecision instead of the negation of the stronger term). Semantic similarity in turn is expected to depend on, among other things, the precise senses in which the scalar terms are used, which can be constrained by context. To test these expectations, we analyzed the experimental results from van Tiel et al. (2016) by fitting models based on different notions of semantic similarity: context-insensitive (LSA, Word2Vec, GloVe) and context-sensitive (ELMo).

Our interpretation of the results supports three main conclusions. First, a sufficiently fine-grained notion of semantic similarity indeed affects scalar inference, and (for our best models) this effect is negative, suggesting that distinctness but not availability may have a role to play in van Tiel et al.'s experiments. Second, it appears that context-sensitivity can improve model performance, but only modestly, only when the stimuli contain informative context words (Experiment 2), and only when the scalar terms themselves are sufficiently context-sensitive (open-class items). Third, even our best models explain only around 6% (Experiment 1) and 4% (Experiment 2) of the variance in the data. The latter may reflect that even our context-sensitive models failed to take an important aspect of the context into account, namely the experimental setting itself. But it is also likely that factors other than semantic similarity play a role. In the future we hope to extend an analysis like the present one to more data and based on a more complete model of scalar inference (e.g., Gotzner et al., 2018).

References

Aina, L., K. Gulordava, and G. Boleda (2019). Putting words in context: Lstm language models and lexical ambiguity. In *Proceedings of the 57th Annual Meeting of the Association*

for Computational Linguistics; 2019 Jul 28-Aug 2; Florence, Italy. Stroudsburg (PA): ACL; 2019. p. 3342–8. ACL (Association for Computational Linguistics).

Baroni, M., G. Dinu, and G. Kruszewski (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Volume 1, pp. 238–247.

Boleda, G. (2020). Distributional semantics and linguistic theory. *Annual Review of Linguistics*.

Clark, S. (2015). Vector space models of lexical meaning. *The Handbook of Contemporary semantic theory*, 493–522.

Doran, R., R. Baker, Y. McNabb, M. Larson, and G. Ward (2009). On the non-unified nature of scalar implicature: an empirical investigation. *International Review of Pragmatics 1*(2), 211–248.

Geurts, B. (2011). *Quantity Implicatures*. Cambridge University Press.

Gotzner, N., S. Solt, and A. Benz (2018). Scalar diversity, negative strengthening, and adjectival semantics. *Frontiers in psychology 9*, 1659.

Harris, Z. S. (1954). Distributional structure. *Word 10*(2-3), 146–162.

Hochreiter, S. and J. Schmidhuber (1997). Long short-term memory. *Neural computation 9*(8), 1735–1780.

Landauer, T. K. and S. T. Dumais (1997a). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review 104*(2), 211.

Landauer, T. K. and S. T. Dumais (1997b). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review 104*(2), 211.

Lenci, A. (2018). Distributional models of word meaning. *Annual review of Linguistics 4*, 151–171.

McNally, L. (2017). Scalar alternatives and scalar inference involving adjectives: A comment on van tiel, et al. 2016. *Asking the right questions: Essays in honor of Sandra Chung*, 17–28.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*.

Pennington, J., R. Socher, and C. D. Manning (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer (2018). Deep contextualized word representations. In *Proc. of NAACL*.

Schwarz, N. (1996). *Cognition and Communication: Judgmental Biases, Research Methods and the Logic of Conversation*. Hillsdale, NJ: Erlbaum.

Seabold, S. and J. Perktold (2010). statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

Van Tiel, B., E. Van Miltenburg, N. Zevakhina, and B. Geurts (2016). Scalar diversity. *Journal of Semantics 33*(1), 137–175.

Westera, M. (2017). *Exhaustivity and intonation: a unified theory*. Ph. D. thesis, submitted to ILLC, University of Amsterdam.

Westera, M. (2020). Implying or implicating 'not both' in declaratives and interrogatives. To appear in Proceedings of Sinn und Bedeutung 24.