A closer look at scalar diversity using contextualized semantic similarity

Matthijs Westera & Gemma Boleda, Universitat Pompeu Fabra

1. Scalar diversity. Van Tiel et al. (2016) show experimentally that the perceived presence of *scalar inferences* varies greatly, with stimuli like:

> *John says: "This sand is warm". Would you conclude from this that, according to John, the sand is not hot? Yes/No.*

They perform two experiments with 25/30 participants each, each with the same 43 pairs of words, like *warm*/*hot* and *adequate*/*good*. The experiments differ in whether the subject is a pronoun (Exp.1) or a more descriptive noun phrase (Exp.2). Results comprised the full range between 0% and 100% of participants choosing *yes*.

Van Tiel et al. adopt the common assumption that scalar inference derives from reasoning about why the speaker didn't mention a relevant stronger alternative: e.g., the speaker said *warm* because *hot* would have been false. They then consider various factors in search of an explanation for the observed scalar diversity, which we review in the full paper, including Semantic Similarity in distributional semantics, for which they found no significant effect. The latter is surprising, as one would expect it to have one or both of the following effects (noted by Van Tiel et al.): First, two terms are similar in distributional semantics if they occur in the same types of contexts, which is expected of words that are pragmatic alternatives. Second, if two terms are semantically *too* similar it may be hard to distinguish them, and a speaker's choice for the weaker term may be due to imprecision rather than falsity of the stronger term. It seems unlikely that in the stimuli of Van Tiel et al. these two effects of semantic similarity would happen to cancel each other out exactly – so why do they not find an effect?

2. Context. Distributional semantics represents words as high-dimensional numerical vectors that are abstractions over their use in large amounts of data. Such representations have been shown to correlate with many aspects of word meaning (for an overview see Clark 2015). Semantic similarity between two words is computed as the cosine of the angle between their vector representations. Van Tiel et al. obtain their measure of semantic similarity from the implementation of Latent Semantic Analysis (LSA) at `lsa.colorado.edu`. In a commentary on Van Tiel et al., McNally (2017) notes that the representations of standard distributional semantics (like LSA) are purely *lexical*: the same vector is assigned to a word regardless of the sentence in which it occurs. Accordingly, Van Tiel et al.'s measure of semantic similarity is context-insensitive in the same way. But McNally notes that sentential context matters for scalar inference, in at least two ways:

- It guides what the relevant alternatives are likely to be: e.g., *the sand is warm* may imply *not hot* because hot sand can be dangerous hence relevant, unlike *the soup is warm* which potentially contrasts warm soups with *cold* soups.
- It guides interpretation of the scalar terms themselves, e.g., although *good* and *adequate* form a scale, when judging whether *the salary is adequate* implies *the salary is not good* one can interpret *adequate* relative to one standard (e.g., meeting one's needs) and *good* relative to another (e.g., being better off than peers).

Hence a more suitable notion of semantic similarity must take context into account.

In this paper we analyze Van Tiel et al.'s data using a recent model developed in computational linguistics that does exactly that, and show that *context-dependent* semantic similarity indeed does have a significant effect. LSA-style distributional semantics, of the sort used by Van Tiel et al., has in the last decade given way almost entirely to neural-network based approaches (Baroni et al. 2014). Recently these neural networks have become *deep*: they start from lexical (context-invariant) representations and combine

them recursively into *contextualized* word representations. We use the highly successful model ELMo (*Embeddings from Language Models*; Peters et al. 2018).

3. Modeling and discussion We apply ELMo to the sentential stimuli of Van Tiel et al. and extract two kinds of representations for the scalar terms: purely lexical (context-invariant) and contextualized word vectors. Cosine similarity of these vectors gives us lexical and context-dependent measures of semantic similarity, which we term ELMo-Lex and ELMo-Con. As a sanity check we also consider the LSA-based similarities used by Van Tiel et al. We fit separate linear regression models (alpha level: .05), each with one of the similarity measures as independent variable. As dependent variable we use the percentage of *yes* responses from either Exp.1 or Exp.2 reported by Van Tiel et al. (we did not have access to individual judgments per participant).

In line with Van Tiel et al., no LSA-based model reached significance. By contrast, both ELMo-based models show significant negative effects for both Exp.1 and Exp.2 (p-values around .002), with a slightly larger effect for ELMo-Cont. The fact that even the context-invariant ELMo-Lex shows an effect suggests that it is a more accurate model than LSA also without context. The larger effect of ELMo-Con compared to ELMo-Lex is weak confirmation of McNally's hypothesis that context matters. However, these results are influenced by the fact that the 4 closed-class stimuli (*some/all*, *may/will*, *may/have to*, *few/none*) have much lower ELMo-similarities than the open-class stimuli. Removing them (leaving 39 of 43 items) reveals a picture in line with McNally's hypothesis, with no model reaching significance on Exp.1 (where sentential context was *un*informative), and only ELMo-Con reaching significance on Exp.2 (where sentential context was informative). Details of the latter:

ELMo-Lex:	R^2:	.061	β:	-.6402	SE: .413	p-value:	.130
ELMo-Con:		.127		-1.441	.620		.026 *
(LSA:		.008		0.1461	.280		.604)

(Adding Van Tiel et al.'s Distance as a factor doesn't change the results; adding Boundedness shows the same tendency but without significance; and likewise for multiple linear regression based on both ELMo-measures at once.) Our results suggest that semantic similarity does affect scalar inference when taking context into account, as hypothesized by McNally. As for why this effect is negative, we hypothesize that explicitly asking participants about the stronger term renders these salient and relevant quite independently of semantic similarity, such that the only remaining effect of semantic similarity is that scalar inference is blocked if the terms are *too* similar (a possibility noted by Van Tiel et al.).

References: Baroni, M., Dinu, G., & Kruszewski, G. (2014). Dont count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *Proceedings of ACL 52.* • Clark, S. (2015). Vector space models of lexical meaning. *Handbook of Contemporary semantic theory.* • McNally, L. (2017). Scalar alternatives and scalar inference involving adjectives: A comment on Van Tiel, et al. 2016. *Essays in honor of Sandra Chung.* • Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *Proceedings of NAACL.* • Van Tiel, B., Van Miltenburg, E., Zevakhina, N., & Geurts, B. (2016). Scalar diversity. *Journal of Semantics* 33.

Acknowledgments: 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).