# How to determine the role of alternatives in scalar implicatures*

Emiel van Miltenburg
VU University Amsterdam
emiel.van.miltenburg@vu.nl

Draft, January 14, 2015
Comments are very welcome!

## 1 Introduction

When someone utters (1a), we can infer (1b) on the basis of the entailment scale (2a) associated with *some*: the use of *some* implies that the use of any stronger expression is not warranted (Horn 1972 and many others since). For classical scales involving quantifiers, modals, or numerals, it is trivial to construct entailment scales (2a–c).

(1) a. Bill ate some of the eggs.
b. Bill didn't eat many/most/all of the eggs.

(2) a. ⟨some, many, most, all⟩
b. ⟨might, can, must⟩
c. ⟨one, two, three, . . . ⟩

Adjectival scales are different from the above-mentioned classical scales, because (i) the lexical class of adjectives is open, and (ii) the set of adjectives is orders of magnitude larger than e.g. the set of quantifiers. As a consequence, there are many more lexical items that could serve as a scalar alternative to any given adjective. Here the question arises: given an adjective, how can we determine its *alternative set*?[1] In other words: how do we find out which other adjectives the hearer considers as an alternative? An answer to this question enables us to study the inferential process in terms of the size and structure of the alternative set, which might explain the variation in the rate at which adjectives give rise to scalar inferences (Van Tiel et al., 2014).

## 2 Approximating the alternative set

While I do not see a way to definitively establish the alternative set for a particular adjective in a given context, there are ways to *approximate* the alternative set. Most of these are based on the idea of *spreading activation*: whenever someone hears a word, semantically related lexical items are activated in the mind/brain to facilitate retrieval from the lexicon (Collins & Loftus, 1975). It stands to reason that the alternative set may be generated in a similar fashion. As such, measures of semantic relatedness (MOSRs) play a significant role in hypothesizing about the alternative set.

Before discussing different MOSRs, it is important to note that there is a difference between *association* and *similarity*. Hill et al. (2014) provide the following example: while *coffee* is associated to *cup*, they are not similar (they are very different things). By contrast, cars and trains *are* similar (both are

*This work was carried out in the *Understanding language by machines* project, supported by the 2013 NWO Spinoza grant to Piek Vossen. I wish to thank Bob van Tiel and members of the Computational Lexicology and Terminology Lab at the VU University Amsterdam for discussion.

[1] I use the term *alternative set* instead of *entailment scale* in order to emphasize the members of the set rather than the order they are in. How to order a particular set of adjectives is a separate issue that lies outside the scope of this paper.

means of transportation), but they aren't closely associated to each other. Both similarity and association are important in determining the alternative set. To see this, consider the examples in (3). The first example (3a) is a typical scalar inference: *warm* is lower on the entailment scale than the more expressive *hot*, and so we may infer that the water is not hot. The second example provides an inference on the basis of the stereotype that good looks do not go together with intelligence.

(3) a. The water is warm ⇒ it is not hot. (*warm* and *hot* are similar and associated)

b. John is handsome ⇒ he is not intelligent. (*handsome* and *intelligent* are dissimilar but associated)

## 2.1 Psycholinguistic measures

A quick way to find the words that are related to a particular set of words is to carry out a free association task: present participants with a word and ask them to write down any word(s) that come to mind (e.g. Nelson et al. 2004). For *similar* words, one might ask participants to write down (near-) synonyms of the word presented to them. A crude way to quantify relatedness is to count the number of times participants had the same response (Weller & Romney, 1988, ch. 2), but most of the relatedness data nowadays is collected through online questionnaires, where participants have to provide a relatedness-rating on a scale (see Hill et al. 2014 for an overview).

A different way to determine semantic relatedness is to perform a priming task. In a lexical priming task, participants are asked to judge whether words appearing on a screen are actual English words or non-words. It has been shown that participants are faster to respond if the preceding word is semantically related (Neely, 1977). In other words: words prime participants to recognize semantically related words quicker. A variation on this is the cross-modal priming task (Swinney, 1979), where priming doesn't occur through preceding words, but rather participants hear a series of utterances through a pair of headphones. Swinney shows that words related to the utterances are recognized faster. It may be possible to adapt this experiment to find out *online* which words are in the alternative set.

## 2.2 Corpus-based measures

Rather than carrying out labor-intensive experiments, we can also use corpus information to determine relatedness. A common method is to use a vector space model (Turney et al., 2010, VSM). In a VSM, words are represented as feature vectors. And since vectors can be interpreted as points in space, we can measure the distance between words to get an idea of how related they are.[2] The big advantage of VSMs is that after training a VSM on a corpus, you instantly have a distance measure between all the words occurring in the corpus. One of the biggest freely available models is trained using the `word2vec` tool on 100 billion words, and has a vocabulary of 3 million words and phrases (Mikolov et al., 2013).[3] See Baroni et al. (2014) for an indication of the quality of models produced using `word2vec`.

Still in the vector space domain is *Clustering by Committee* (Pantel, 2003, CBC), an algorithm that tries to discover concepts and cluster together words expressing similar meanings. Insofar as thesaurus data is limited in coverage, CBC and other automatic thesaurus generation algorithms show us a wider range of related expressions for any given word.

Another way to approximate the alternative set is to use lexical patterns, such as *A if not B*, to find adjectives that are semantically related. Hatzivassiloglou & McKeown (1993) suggest to use this pattern-based method based on Horn's (1969) observation that patterns like *A even B* can be used to test whether

[2] Usually, the *cosine similarity* measure is used in the VSM literature, which is the cosine of the angle between two vectors.

[3] The model is available at `https://code.google.com/p/word2vec/`. The Gensim module in Python provides an intuitive interface. The code below loads the pre-trained model and prints out a list of words that are related to *warm* and *hot*, but not to *cold*. These are: *hottest, warmed, hot, heated, toasty, hotter, warms, sizzling, scorching, cool*.

```
from gensim.models import Word2Vec
model = Word2Vec.load_word2vec_format('GoogleNews-vectors-negative300.bin', binary=True)
print model.most_similar(positive=['warm','hot'],negative=['cold'])
```

A and B are scalemates. However, Hatzivassiloglou & McKeown did not pursue this endeavor because they lacked a large enough corpus. van Miltenburg (To appear) uses a pattern-based method with a 3 billion-word corpus to generate lists of potential scalemates.[4]

# 3 Conditions and restrictions on the alternative set

So far, we have discussed the alternative set, and how we can understand this set as the result of a proces of spreading activation. We operationalized the notion of spreading activation in terms of semantic relatedness, and suggested different ways of finding words that are related to a given expression. In other words: we have talked about how to fill and expand the alternative set. But there are also other forces that manipulate the alternative set. I discuss some of these below.

## 3.1 Distance

While the alternatives to a given expression should generally be related, they should not be *too* close to each other. For example, *beautiful* and *gorgeous* are equivalent (save for their selection restrictions). When a speaker utters one or the two, we can infer that the other is also the case and so they are not in each other's alternative set. Van Tiel et al. (2014) show that a greater distance between two lexical items on the same scale increases the chance of eliciting an implicature. They assessed the distance between scalar expressions by using an online questionnaire, asking participants to rate the difference in 'strength' on a 5-point likert scale. De Marneffe et al. (2010) and Potts (2011) present some ways to automatically determine the distance between two scalar expressions. Sentiment-related expressions can be compared on the basis of review data, and how strongly the use of these expressions correlates with the positivity or negativity of the reviews. Some other expressions, such as those denoting temperatures, can be compared on the basis of numerical information that usually accompanies them. For example, *warm* might commonly occur with temperature values of around 20–28 °C, while *hot* occurs with higher temperature values, e.g. 28–38°C. I do not know of a general computational measure of strength differences between two related adjectives (Potts' review based approach only covers sentiment-related adjectives, and De Marneffe et al.'s approach covers adjectives that may be numerically quantified), but feel that such a measure would make a valuable addition to the field.

## 3.2 The question under discussion

Expressions in the alternative set should be relevant to the *Question under Discussion* (Roberts, 2012). As Djalali et al. (2011) note, "any discourse can be viewed as a sequence of questions and their answers, all of which address sub-issues of the current topic of conversation, until all such issues are exhaustively resolved." It is difficult to see how, given a particular context, one might automatically identify all questions under discussion, but it *is* possible to identify topics. Latent Dirichlet Allocation (Blei et al., 2003, LDA) is a method commonly used for document classification. It rests on the assumption that each document contains a mixture of different topics. Topics are modeled as probability distributions over a vocabulary. LDA classifies documents by comparing the distribution of words in each document to the probability distributions of the topics it is trained on. This also works the other way round: given a particular topic, the model predicts which words are likely to occur. So if we have full knowledge of the topics in a particular interaction, we also know which words we can expect to be used.

## 3.3 Other factors

There are many other factors to discuss in this short paper. For example, alternatives should be congruent with the subject: the adjective *handsome* is only used for males while *pretty* is used for females. When

[4]The lists are available through `http://kyoto.let.vu.nl/~miltenburg/public_data/adjectival-scales/output/potentials/`

the subject of the conversation is a handsome male, *pretty* cannot be in the alternative set, even though it *is* a related property. Van Tiel (2012) provides a good overview of the factors involved, including register (*transcendent* vs. *great*), granularity (*fifty* versus *many*), frequency, complexity (*spine-chilling* vs. *scary*) and monotonicity (*sometimes* vs. *intermittently*). Most of these are fairly straightforward to operationalize, with frequency being the most trivial case. Register may be determined through a corpus study, comparing texts from different sources. Complexity is likely to correlate with frequency and word length (cf. concreteness, Feng et al. 2011), and otherwise one might use morphological software.[5] Monotonicity and granularity seem to be the toughest nuts to crack.

# 4 Conclusion

With this contribution, I have outlined the problem of determining the alternative set for a given utterance. In other words: determining what the speaker also could have said, but didn't. There are two general processes determining this set of expressions: (i) the process of associating lexical items with the utterance, and (ii) using the context to restrict or modify the set of possible alternatives. I have tried to operationalize these processes in terms of psycholinguistic experiments and natural language processing (with some emphasis on the latter). While most of the individual factors are relatively easy to model or approximate, the question remains what the general alternative generation architecture should look like. Nevertheless, I hope to have provided a handle on the issue so as to make further research possible. I believe that it is vital for researchers in pragmatics and computational linguistics to work together, especially as pragmatics is becoming more and more data-driven. For computational linguists, it is important to diversify the kinds of data used to evaluate their models on. These are fruitful grounds for collaboration.

# References


Baroni, Marco, Georgiana Dinu & Germán 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*, vol. 1, .

Blei, David M, Andrew Y Ng & Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3. 993–1022.

Collins, Allan M & Elizabeth F Loftus. 1975. A spreading-activation theory of semantic processing. *Psychological review* 82(6). 407.

Djalali, Alex, David Clausen, Sven Lauer, Karl Schultz & Christopher Potts. 2011. Modeling expert effects and common ground using questions under discussion. In *Aaai fall symposium: Building representations of common ground with intelligent agents*, .

Feng, Shi, Zhiqiang Cai, Scott A Crossley & Danielle S McNamara. 2011. Simulating human ratings on word concreteness. In R. Charles Murray & Philip M. McCarthy (eds.), *Proceedings of the twenty-fourth international florida artificial intelligence research society conference*, AAAI Press.

Hatzivassiloglou, Vasileios & Kathleen R McKeown. 1993. Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning. In *Proceedings of the 31st annual meeting on association for computational linguistics*, 172–182. Association for Computational Linguistics.

Hill, Felix, Roi Reichart & Anna Korhonen. 2014. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *arXiv preprint arXiv:1408.3456* .

Horn, Laurence. 1969. *A presuppositional analysis of only and even*. RI Binnick.


[5] See `http://aclweb.org/aclwiki/index.php?title=Morphology_software_for_English`.

Horn, Laurence. 1972. *On the semantic properties of logical operators in english*: University of California, Los Angeles dissertation.

de Marneffe, Marie-Catherine, Christopher D Manning & Christopher Potts. 2010. Was it good? it was provocative. learning the meaning of scalar adjectives. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, 167–176. Association for Computational Linguistics.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado & Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

van Miltenburg, Emiel. To appear. Detecting and ordering adjectival scalemates. Paper to be presented at MAPLEX 2015.

Neely, James H. 1977. Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of experimental psychology: general* 106(3). 226.

Nelson, Douglas L, Cathy L McEvoy & Thomas A Schreiber. 2004. The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers* 36(3). 402–407.

Pantel, Patrick Andre. 2003. *Clustering by committee*: University of Alberta dissertation.

Potts, Christopher. 2011. Developing adjective scales from user-supplied textual metadata. NSF Workshop on Restructuring Adjectives in WordNet. Arlington,VA.

Roberts, Craige. 2012. Information structure: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics* 5(6). 1–69. doi:10.3765/sp.5.6.

Swinney, David A. 1979. Lexical access during sentence comprehension:(re) consideration of context effects. *Journal of verbal learning and verbal behavior* 18(6). 645–659.

van Tiel, Bob. 2012. Scalar alternatives. Unpublished manuscript.

Turney, Peter D, Patrick Pantel et al. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research* 37(1). 141–188.

Van Tiel, Bob, Emiel Van Miltenburg, Natalia Zevakhina & Bart Geurts. 2014. Scalar diversity. *Journal of Semantics* doi:10.1093/jos/ffu017. First published online: December 23, 2014.

Weller, Susan C & A Kimball Romney. 1988. *Systematic data collection*, vol. 10. Sage.