

Notes from 1st brainstorming

Barbara and Michael

February 8 2018

1 Background: theories and experimental data

2 Models

Rather than redo, outwit or refute previous experimental contributions, **our goal is to chart new territory**. The starting point should be new predictions made by conceptually interesting probabilistic models, ideally extensions of the optimal- θ model or the RSA model for GASs (Lassiter and Goodman, online first; Qing and Franke, 2014a,b). Two ideas came to mind.

2.1 Lexical uncertainty about absolute GASs

Lexical uncertainty models (Bergen, Levy, and Goodman, 2012, to appear; Potts et al., 2016) assume that the listener is uncertain about the lexical meaning that a speaker might bring to the conversation. We consider uncertainty about the lexical meaning of absolute GASs: do they receive a relative (prior dependent) or absolute (pure standard) interpretation. We combine lexical uncertainty with the GA-model of Lassiter and Goodman (online first) (exactly what Tessler and Franke, 2018, did too):

$$L_1(x, \theta, \mathcal{L} \mid u) \propto S_1(u \mid x, \theta, \mathcal{L}) \cdot P(x) \cdot P(\theta) \cdot P(\mathcal{L}) \quad (1)$$

$$S_1(u \mid x, \theta, \mathcal{L}) \propto \exp(\alpha \cdot \ln L_0(x \mid u, \theta, \mathcal{L}) - \text{cost}(u)) \quad (2)$$

$$L_0(x \mid u, \theta, \mathcal{L}) \propto \mathcal{L}(u, x, \theta) \cdot P(x) \quad (3)$$

A lexicon is a map $\mathcal{L}: u, x, \theta \mapsto \{0; 1\}$ which gives a (Boolean) truth-value for any utterance u of some GA, degree x and threshold θ . Only absolute gradable adjectives are lexically uncertain in the way described above. Model variants could distinguish cases where speakers maintain a single rule (all absolute GASs are prior-dependent/pure-standard) or between-item flexibility (e.g., *full* is prior dependent; *bent* is not).

This model is likely to make interesting **novel predictions about task effects** that other stories are unlikely to offer anything beyond hand-wavy explanations. Generally speaking, observations from previous trials/encounters could shift beliefs about the speaker's likely lexicon. If speaker's have been observed to use an absolute GA to refer to non-absolute degree x , listeners should update their lexical beliefs accordingly and be more likely to interpret a future use of this GA (or others, depending on the model variant) as relative-standard (prior-dependent). Also, interpretation tasks which display multiple utterances at the same time ((implicitly:) by the same speaker) could show interesting effects of jointly conditioning the model with all observed utterances (as observed by Tessler and Franke, 2018).

[mf: models need to be formulated precisely, implemented and predictions checked; this is all just intuitive guesses about potential model predictions]

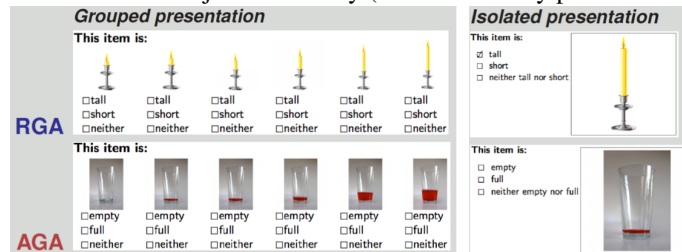
2.2 Uncertainty about the prior (or the comparison class)

If speakers and listeners are uncertain about the prior over degrees $P(x)$, we are also bound to see potentially interesting **predictions about response dynamics as a function of prior exposure**. Suppose that items to be judged or chosen for interpretation are presented individually in each trial, or at the same time (like in stuff from the Chicago group (Kim, Xiang, and Kennedy, 2014; Leffel, Xiang, and Kennedy, 2016) [mf: insert ref]), this task manipulation will likely have effects on participants' construction of the comparison class / the relevant prior distribution. For example, for absolute GASs it might matter whether the end-point degrees have already been observed or not: as long as there

is uncertainty about how likely these belong to the comparison class, priors with little density on these degrees are reasonably likely, thus shifting predictions about θ “further away from the end-points”. In general, the more extreme instances are observed, the more median instances should count as “neither this nor that”.

2.3 Relevant experimental results

Kim, Xiang, and Kennedy (2014) find evidence that absolute, but not relative, adjectives show signs of having context-invariant, precise meanings. In Experiment 1, they find that the thresholds for relative adjectives are sensitive to the local context (grouped vs. isolated presentation in the picture below), but those for absolute adjectives are not. They find a significant effect of scale position for relative adjectives in grouped presentation (but not in isolated presentation); and no difference for absolute adjectives by presentation type. In the judgments for relative adjectives, there was more variance due to noun identity (in isolated presentation). In the judgments for absolute adjectives, there was more variance due to adjective identity (no difference by presentation type).



In Experiment 2 with just isolated presentation, they find that only absolute adjectives show asymmetric shiftability: they allow shifts to higher levels of precision, and resist loosening a standard that was previously set at maximum precision (main effect of previous precise exemplar (i.e. previous exposure to items like the completely empty cup above) and an interaction with scale position, but no interaction with number of intervening uses of the same adjective). In contrast, relative adjectives allow loosening subsequent standards irrespective of scale position, and the effect decreases with the number of intervening uses (symmetric shiftability).

In Experiment 3 (isolated presentation), they find that both adjective types are sensitive to contexts where goals supporting higher/lower standards are explicitly introduced, e.g., *tall ladder* to get a kite stuck on a chimney vs. to get a book from the top shelf; *empty cup* to play a game of flip-cup vs. to be refilled at a party. However, when the contexts make standards irrelevant, only absolute adjectives revert to high-precision standards.

The same design is used in Leffel, Xiang, and Kennedy (2016) with an additional manipulation: type of object. They used (a) pictures of familiar, everyday objects (about which we can assume rich prior knowledge about degree distribution within the comparison class) and (b) artificially-constructed images of geometric shapes like cubes, pentagrams, and arrows. They found no effect of presentation type (grouped vs. isolated), but main effects of scale position and adjective type, and interactions: object type * adjective type, object type * adjective type * scale position. With shape nouns there were more end-point oriented interpretations than with artifacts. (For artifacts, similar results were obtained by Foppolo and Panzeri. (2011).) In a post-hoc test, they also find that previous exposure to maximal exemplar is a reliable predictor of response for absolute adjectives (but not relative).

3 Materials and envisaged pilot

With sufficient background knowledge, is there a qualitative difference in the process of setting the threshold for absolute adjectives vs. relative adjectives – is there a difference in the RTs and ERPs?

Background knowledge: adjectives are presented with familiar artifacts (for which comparison classes are familiar).

Language: German

14 absolute, 14 relative adjectives; 6 artifacts for each adjective (168 adjective-artifact pairs)

Stimulus sentence: *Dieses Objekt ist...*

- *absolute maximum standard*: trocken (dry), voll (full), leer (empty), glatt (smooth, even, plain), klar (clear), gerade (straight), geschlossen (closed), sauber (clean)
- *absolute minimum standard*: nass (wet), gepunktet (spotted), trüb (dull, dim, foggy, cloudy), gebogen (bent, curved), offen (open), dreckig (messy, dirty)

- *relative*: lang (long), kurz (short), gross (big - all dimensions, including 'tall'), klein (small), dick (fat, thick), dnn (thin), hell (bright, fair), dunkel (dark), schwer (heavy), leicht (light), hoch (high), niedrig (low), scharf (sharp), stumpf (dull, blunt)

3.1 Experiment 0 - Eliciting the priors

Schöller and Franke (2017) obtain an empirical estimate of participants prior expectations for *many/few* by asking participants for numerical values appropriate in a particular context. E.g. *Andy is a man from the US. How many cups of coffee do you think Andy drank last week?*. In the elicitation questions the quantifiers *many/few* are not used. ⇒ Is it possible to do something like this for *dry facemask, wet sunglasses, open door*, etc. where the properties cannot be mapped onto a numerical scale?

3.2 Experiment 1 - Rating task on a 5-point scale






Scale position 1: minimal exemplar






Scale position 5: maximal exemplar

Between-subjects factor: presentation type (groups of 5 as in the examples below vs. isolated, i.e., one object per screen)





Absolute Adjectives World knowledge allows for imprecise standards. Hypothetical judgments:






Maximum standard

				
This is straight. <input type="checkbox"/> Yes. <input checked="" type="checkbox"/> No.	This is straight. <input type="checkbox"/> Yes. <input checked="" type="checkbox"/> No.	This is straight. <input type="checkbox"/> Yes. <input checked="" type="checkbox"/> No.	This is straight. <input checked="" type="checkbox"/> Yes. <input type="checkbox"/> No.	This is straight. <input checked="" type="checkbox"/> Yes. <input type="checkbox"/> No.

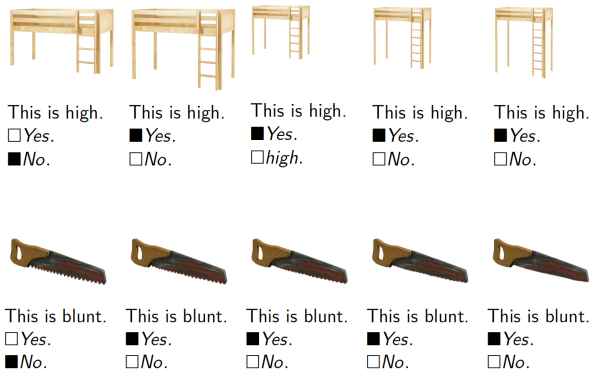
				
This is straight. <input type="checkbox"/> Yes. <input checked="" type="checkbox"/> No.	This is straight. <input type="checkbox"/> Yes. <input checked="" type="checkbox"/> No.	This is straight. <input type="checkbox"/> Yes. <input checked="" type="checkbox"/> No.	This is straight. <input type="checkbox"/> Yes. <input checked="" type="checkbox"/> No.	This is straight. <input checked="" type="checkbox"/> Yes. <input type="checkbox"/> No.

Minimum standard

				
This is bent. <input type="checkbox"/> Yes. <input checked="" type="checkbox"/> No.	This is bent. <input type="checkbox"/> Yes. <input checked="" type="checkbox"/> No.	This is bent. <input checked="" type="checkbox"/> Yes. <input type="checkbox"/> No.	This is bent. <input checked="" type="checkbox"/> Yes. <input type="checkbox"/> No.	This is bent. <input checked="" type="checkbox"/> Yes. <input type="checkbox"/> No.

				
This is bent. <input type="checkbox"/> Yes. <input checked="" type="checkbox"/> No.	This is bent. <input type="checkbox"/> Yes. <input checked="" type="checkbox"/> No.	This is bent. <input checked="" type="checkbox"/> Yes. <input type="checkbox"/> No.	This is bent. <input checked="" type="checkbox"/> Yes. <input type="checkbox"/> No.	This is bent. <input checked="" type="checkbox"/> Yes. <input type="checkbox"/> No.

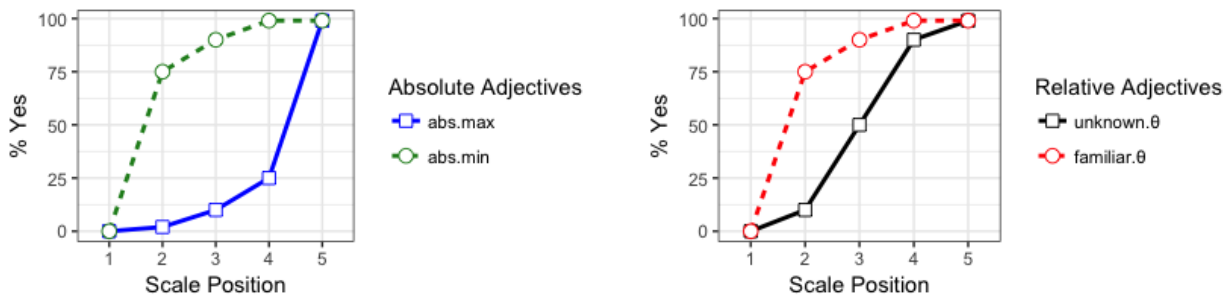
Relative Adjectives World knowledge allows for using scale position 1 as the threshold, e.g. a 'high bed' is already high (sorry! the pics here are not to scale!). Hypothetical judgments:



Hypothetical results

Maximum standard absolute adjectives should elicit some Yes answers already at scale position 3. Conversely, with **minimum standard adjectives** we should get No's also at position 2 (see graph below).

In the experiments by the Chicago group and others, judgments for **relative adjectives unfamiliar thresholds** align in the characteristic S-shape - black line in the graph below. That's because the threshold gets set with respect to the five objects in the comparison class: it's above the 'average', i.e. position 3. **With familiar comparison classes**, we expect that the distribution will shift and resemble the profile for minimum standard adjectives (see also theoretical arguments in e.g., Burnett ...)



3.3 Experiment 2 - Reaction Times

If the ratings profiles for absolute minimum and relative adjectives come out similar, we compare the RTs for the decisions at the corresponding scale positions. ⇒ Experiment 3: compare ERPs.

4 Future music

Two **big issues** to ponder:

- how to derive predictions about response reaction times from prob-models?
- how to link model predictions to data from some suitable EEG study?

References

- Bergen, Leon, Roger Levy, and Noah D. Goodman (2012). “That’s what she (could have) said: How alternative utterances affect language use”. In: *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*.
- (to appear). “Pragmatic Reasoning through Semantic Inference”. to appear in *Semantics & Pragmatics*.
- Foppolo, Francesca and Francesca Panzeri. (2011). “When straight means relatively straight and big means absolutely big”. Paper presented at the 31st Incontro di Grammatica Generativa, Rome.
- Kim, Christina, Ming Xiang, and Chris Kennedy (2014). “Shiftability and goal-dependence in gradable adjectives”. Poster presented at the 88th Annual Meeting of the Linguistic Society of America, Minneapolis.
- Lassiter, Daniel and Noah D. Goodman (online first). “Adjectival vagueness in a Bayesian model of interpretation”. In: *Synthese*.
- Leffel, Timothy, Ming Xiang, and Chris Kennedy (2016). “Context and world knowledge in gradable adjective interpretation”. Talk at the 90th Annual Meeting of the Linguistic Society of America, Washington, DC.
- Potts, Christopher et al. (2016). “Embedded implicatures as pragmatic inferences under compositional lexical uncertainty”. In: *Journal of Semantics* 33.4, pp. 755–802.
- Qing, Ciyang and Michael Franke (2014a). “Gradable Adjectives, Vagueness, and Optimal Language Use: A Speaker-Oriented Model”. In: *Proceedings of SALT 44*. Ed. by Jessi Grieser et al. elanguage.net, pp. 23–41.
- (2014b). “Meaning and Use of Gradable Adjectives: Formal Modeling Meets Empirical Data”. In: *Proceedings of CogSci*. Ed. by Paul Bello et al. Austin, TX: Cognitive Science Society, pp. 1204–1209.
- Schöller, Anthea and Michael Franke (2017). “Semantic values as latent parameters: Testing a fixed threshold hypothesis for cardinal readings of *few* & *many*”. In: *Linguistic Vanguard* 3.1.
- Tessler, Michael Henry and Michael Franke (2018). “Not unreasonable: Carving vague dimensions with contraries and contradictions”. Manuscript.