Tying the denotation of "tall" to considerations of utility

WIP-LANG, ENS

Émile Enguehard

27 November 2018

1 Introduction: gradable adjectives

Gradable adjectives are generally assumed to have a denotation of this sort:

(1) $[\![\text{tall}]\!] = \{x | \, \text{height}(x) \geq \theta\}$
"John is tall" $\approx$ height(John) $\geq \theta$

θ is *vague*: there is no intuitively accessible or experimentally visible limit between tall and non-tall people.

θ may be thought to be a free variable that is supplied (though not fully specified) by context, or to depend in a systematic way on a *comparison class*. In a general way, the question of how θ is determined is still open.

2 Game-theoretic pragmatics and gradable adjectives

There is a recent literature that attempts to link gradable adjectives to game-theoretic / probabilistic models of pragmatics:

1. Lassiter and Goodman (2017) use the RSA model to show that, even if nobody knows anything about θ *a priori*, gradable adjectives may still be used in an informative way by Bayesian reasoners.

2. Qing and Franke (2014) propose that speakers might actually derive expectations about θ from considerations of optimal usage: speakers expect θ to be used so as to maximize a certain function, the *expected utility*.

The idea is to derive non-trivial semantics "for free" from a general-purpose model.

In both cases there is and assumption that there is a *prior distribution* on the quantity being discussed in the common ground.

3 Utility and expected utility

We're going to take the utility of Lassiter and Goodman (2017), apply to it the methodology of Qing and Franke (2014), and derive from that a very clear version of the sort of things these models predict.

1. Utility is a numerical model of a trade-off between being *informativeness* and cost. What's important to us is that it quantifies informativeness.

2. We assume there are two messages, POS which denotes $[\theta, +\infty)$ and has cost c (for instance, "John is tall"), and ε (the empty message) which provides no information and has cost 0.

3. For a given situation, for instance a person being of height h_0, the utility function maps each message to a numerical value. We write:

(2) a. $U(\varepsilon|H = h_0; \theta)$: utility of saying ε for a given value of θ and for height h_0.
b. $U(\textsc{pos}|H = h_0; \theta)$: utility of saying POS for a given value of θ and for height h_0.

Once we have a utility, we may define an *expected* utility: this quantifies how much utility speakers achieve *on average*:

(3) $EU(\theta) = \mathbb{E}_H[U(u^*|H; \theta)]$

We assume speakers are as informative as they can: they say POS whenever it is true.

4 Result

It turns out there is an optimal value for θ, which we write as θ^* such that expected utility is maximal. This value can be expressed in terms of $\Phi(\theta)$, the prior probability of not being tall:

(4) $\Phi(\theta^*) = 1 - \exp(-1 - c)$

Thus, optimally, **gradable adjectives should denote a fixed share of the comparison class**, specifically $\exp(-1 - c)$.

Note that if $c \approx 0$, then $1 - \Phi(\theta^*) \approx \frac{1}{\mathrm{e}} \approx 0.37$. Finally:

(5) Theorem: the optimal denotation of "tall" is such that 37% of the people are tall.

5 Extension

What if we also have the following messages?

$$\begin{aligned}
[\![\textsc{pos}]\!][x] &= H_x \geq \theta_p \\
[\![\textsc{neg}]\!][x] &= H_x \leq \theta_n \\
[\![\neg\textsc{pos}]\!][x] &= H_x < \theta_p \\
[\![\neg\textsc{neg}]\!][x] &= H_x > \theta_n
\end{aligned}$$

If POS is "tall", NEG could be "short", and ¬POS "not tall".

We derive a similar result:

(6) Second theorem: optimally, the bottom 22% of the distribution are short, the top 41% are tall, the in-between people may be referred to as "not tall". "not short" isn't used.

6 Discussion

- The idea that gradable adjectives denote a certain fixed share of the population is already discussed by Schöller and Franke (2015), who call it CFK semantics. It seems to give us vagueness (through underspecfication of the distribution) and cognitive plausibility (there isn't an arbitrary threshold for every comparison class, context-dependency is kept in check).
- The actual numbers are plausible enough relative to intuitions about who is tall. Schöller and Franke (2015) ask subjects for intuitions on "many" and "few" and find thresholds of 0.69 (against 0.59 for us) and 0.15 (against 0.22 for us) respectively.

- However, existing linguistic work on gradable adjective usage doesn't really support these sort of probabilistic approaches:

   a. When you know the distribution perfectly, there appears to be an element of "standing out" (Kennedy 2007) which isn't affected by the distribution, and the vagueness doesn't go away.

   b. Though Qing and Franke (2014) claim it does, it's not clear that the difference between "absolute" gradable adjectives like "full" and "relative" ones like "tall" really follows from these semantics.

- What this work shows is that utility-based models are actually interpretable in qualitative ways if one takes care to analyse them in full.

References

Kennedy, Christopher (2007). 'Vagueness and grammar: The semantics of relative and absolute gradable adjectives'. In: *Linguistics and philosophy* 30.1, pp. 1–45.

Lassiter, Daniel and Noah D. Goodman (Oct. 2017). 'Adjectival vagueness in a Bayesian model of interpretation'. In: *Synthese* 194.10, pp. 3801–3836.

Qing, Ciyang and Michael Franke (2014). 'Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model'. In: *Semantics and Linguistic Theory*. Vol. 24, pp. 23–41.

Schöller, Anthea and Michael Franke (2015). 'Semantic values as latent parameters: Surprising few & many'. In: *Semantics and linguistic theory*. Vol. 25, pp. 143–162.

A Details of the model

a. Possible worlds correspond to a range of values for a real variable H (for instance, it could be John's height).

b. There is a prior distribution on H, with density ϕ and probability function Φ, which is part of the common ground (for instance the distribution of men of John's age's height in John's country).

c. There are two messages to describe H, POS which denotes $[\theta, +\infty)$ and has cost c (for instance, "John is tall"), and ε (the empty message) which provides no information and has cost 0.

d. The *utility* of using a message to describe an instance of H, knowing H and θ, is the log of the posterior density an hypothetical listener would assign to the actual value of H (informativeness term), minus the cost of the message (cost term).

(7)
$$U(\varepsilon|H = h_0; \theta) = \log\phi(h_0|H \in [\![\varepsilon]\!]; \theta) = \log\phi(h_0)$$
$$U(\textsc{pos}|H = h_0; \theta) = \log\phi(h_0|H \in [\![\textsc{pos}]\!]; \theta) - c = \begin{cases} -\infty & \text{if } h_0 < \theta, \\ \underbrace{\log\dfrac{\phi(h_0)}{1-\Phi(\theta)}}_{\text{informativeness}} \underbrace{-c}_{\text{cost}} & \text{if } h_0 \geq \theta. \end{cases}$$

B Derivations

Speakers maximize utility by saying POS if it is true, ε otherwise.[1]

[1] In theory it could be the case that if c is too high and θ is too low, it isn't always worth it to say POS. We will ignore this case.

$$(8) \qquad EU(\theta) = \mathbb{E}_\phi[U(u^*|H;\theta)] = \underbrace{\int_{-\infty}^{\theta} \phi(h)U(\varepsilon|H=h;\theta)\mathrm{d}h}_{\text{utility from saying } \varepsilon} + \underbrace{\int_{\theta}^{+\infty} \phi(h)U(\textsc{pos}|H=h;\theta)\mathrm{d}h}_{\text{utility from saying } \textsc{pos}}.$$

The expected utility turns out to have a very simple form:

$$(9) \qquad \begin{aligned} EU(\theta) &= \int_{-\infty}^{\theta} \phi(h)\log\phi(h)\mathrm{d}h + \int_{\theta}^{+\infty} \phi(h)(\log\frac{\phi(h)}{1-\Phi(\theta)} - c)\mathrm{d}h \\ &= \int_{-\infty}^{+\infty} \phi(h)\log\phi(h)\mathrm{d}h - (1-\Phi(\theta))(\log(1-\Phi(\theta)) + c) \\ &= \underbrace{-H(\phi)}_{\text{constant term}} \underbrace{-(1-\Phi(\theta))(\log(1-\Phi(\theta)) + c)}_{\text{variable term } (f(\Phi(\theta)))}, \end{aligned}$$

The first term, $H(\phi)$, is the *entropy* of the distribution and doesn't depend on θ. The second term is a simple function of $\Phi(\theta)$ (cf. Figure 1). Note that $\Phi(\theta)$ is the share of people who are not tall.

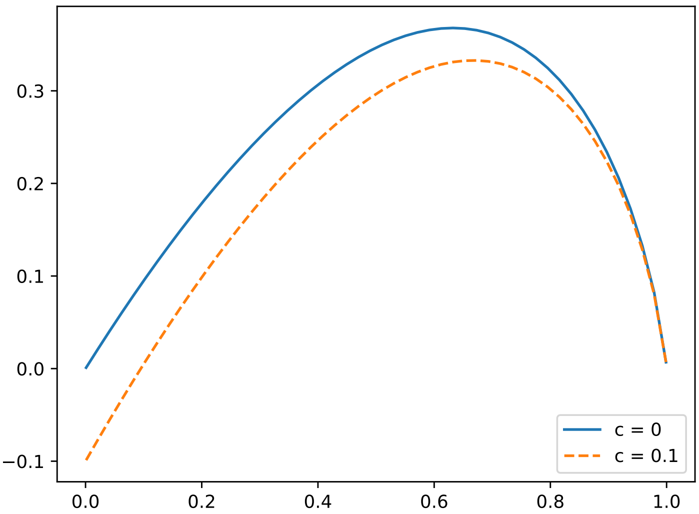

Figure 1: $f(\Phi(\theta))$ as a function of $\Phi(\theta)$ for $c = 0$ and $c = 0.1$.

As can be seen on Figure 1, there is an optimal value for $\Phi(\theta)$ and therefore θ, which we may in fact calculate (through derivatives):

$$(10) \qquad \theta^* = \Phi^{-1}(1 - \exp(-1-c))$$

Exercise: verify[2] that if we introduce the other messages to the model, and we assume that NEG has slightly higher cost than POS, then the optimal system is such that:

$$(11) \qquad \Phi(\theta_p^*) = \frac{1}{1 + \exp(\mathrm{cost}(\neg\textsc{pos}) - \mathrm{cost}(\textsc{pos}) - \mathrm{e}^{-1-\mathrm{cost}(\textsc{neg})+\mathrm{cost}(\neg\textsc{pos})})} \qquad \left(\approx \frac{1}{1+\mathrm{e}^{-\frac{1}{e}}} \approx 0.59\right)$$

$$\Phi(\theta_n^*) = \exp(-1 - \mathrm{cost}(\textsc{neg}) + \mathrm{cost}(\neg\textsc{pos}))\Phi(\theta_p^*) \qquad \left(\approx \frac{1}{\mathrm{e}(1+\mathrm{e}^{-\frac{1}{e}})} \approx 0.22\right)$$

[2] If it turns out to not be true, do tell me.