Evaluativity: a proposal for empirical investigation

Floris T. van Vugt
florisvanvugt@ucla.edu

June 7, 2010

1 Markedness

Adjectives that denote into scales, such as *long*/*short*, often come in pairs that are asymmetric. For example, Clark (1969) observes that only one can denote the scale itself (e.g. *length* but **shortness*), and only one can combine with a measure phrase, as in (1). Traditionally, the adjective that has the more restricted distribution (e.g. *short*) has been called *marked*, and its contrary *unmarked*. Similarly, it is felt that when the marked term is used in questions this is understood as presupposing that it applies to the argument. For example, asking (2-b) seems to presuppose that John is short, which contrasts with (2-a) which seems to presuppose nothing about John, not even that he would be tall. Finally, it is reported that marked terms are learned later during child language acquisition(Clark, 1972).

(1) a. John is 5ft tall.

b. *John is 5ft short.

(2) a. How tall is John?

b. How short is John?

In the seventies a number of experimental studies was performed that constitute evidence for the psychological reality of the marked–unmarked distinction (see for example Seymour (1974)). Chase & Clark (1971) investigated the marked/unmarked pair *below*/*above* and report that subjects had more difficulty affirming that a star was below the circle than that the circle was above the star. The salient explanation is that *below* denotes a concept that is encoded in a somehow more complex way than *above* (for a more detailed discussion of these sentence–picture verification tasks, see Chase & Clark (1972)).

In a recent criticism, Proctor & Cho (2006) suggests that these experimental results can be explained in a more general framework, in which the advantage to affirming the *above*–sentences relative to affirming the *below*–sentences stems from the fact that the affirmative response itself has some abstract positive polarity. This positive polarity aligns with the stipulated polarity *above* but will produce a mismatch with that of *below*, causing the reaction time differences (a version of this idea has also been presented in Carpenter & Just (1975)). Though a detailed survey of this discussion is beyond the scope of this paper, it is important to note that the reaction time difference that was reported in the seventies may not be due to processing

difficulty of marked terms in themselves, but rather their interplay with the required response.

1.1 Markedness in comparatives

Higgins (1977) reported that comparatives with marked terms are more presuppositional than those with unmarked terms. Presuppositionality is understood as follows. When someone utters (3-a), in some accounts this presupposes that both Bob and Fred are bad. However, (3-b) does not seem to presuppose either Bob and Fred being bad or good.

(3) a. Bob is worse than Fred

b. Bob is better than Fred

c. Fred is better than Bob

d. Fred is worse than Bob

Higgins (1977) measured the presuppositionality in an *acceptability* task, where subjects were asked to rate the acceptability of a sentence that compared two items that clearly had a quality opposite to the one implied by the adjective. Such sentences with a marked adjective, e.g. (4-b), were judged less acceptable than those with an unmarked one, e.g. (4-a).

(4) a. A feather is heavier than a snowflake

b. A mountain is lighter than a ship

The author argues that both results can be explained by the marked adjectives carrying the presupposition that the entities that are compared possess the marked quality. For example, (3-a) implies that both Bob and Fred are bad, but (3-c) does not imply that they are good, hence they are perceived as less synonymous. Similarly, the use of the marked adjective in (4-b) implies that the arguments are light, which is not the case, causing subjects to perceive the sentence as less acceptable.

I would argue, however, that the fact that marked adjectives are much less frequent than unmarked adjectives caused participants to relatively disprefer a sentence with a marked adjective. In order to control for this, it would have been desirable to compare ratings of sentences in (5). If the acceptability difference is absent here, this would constitute evidence that it is due to the adjective markedness and not some other factor.[1]

(5) a. A guilder is heavier than a dollar.

b. A guilder is lighter than a dollar.

2 Evaluativity

In more modern semantic terminology the effect reported in section 1.1 is referred to as *evaluativity*. A phrase is evaluative if "it makes reference to a degree that exceeds a contextually specified standard"(Rett, 2008a). For

[1]When I propose use of Higgins (1977)'s experimental paradigm I will assume that these appropriate controls are performed as well.

example, uttering (6-a) establishes that Boris exceeds a contextually specified standard of tallness. However, (6-b) implies no such thing, and similarly (6-c). (6-d) is again commonly perceived as implying that the individuals that are mentioned are short.

(6) a. Boris is tall.

b. Boris is taller than Doris.

c. Boris is as tall as Doris.

d. Boris is as short as Doris.

Rett (2008b) suggests that markedness plays a role in the evaluativity of comparatives and equatives. In particular, she argues that comparatives are not generally evaluative, regardless of whether marked or unmarked terms are used. This property is referred to as *polarity–invariance*. The equative construction with a marked adjective, however, is usually perceived as evaluative (e.g. (6-d)), and hence the equative is *polarity–variant*.

2.1 Evaluativity in the equative

How can this be explained? Let us first consider the equative. (6-c) could be construed to be ambiguous between (7-a) and (7-b). Now (6-d) can be interpreted analogously by (7-c) or (7-d).

(7) a. $\exists d\ \textsc{max}\{d|\textsc{tall}(\text{Boris}, d)\} = \textsc{max}\{d|\textsc{tall}(\text{Doris}, d)\}$.

b. $\exists d\ \textsc{max}\{d|\textsc{tall}(\text{Boris}, d)\} = \textsc{max}\{d|\textsc{tall}(\text{Doris}, d)\} > d_{\text{tall}}$ for

some contextually specified standard d_{tall}.

c. $\exists d\ \text{MAX}\{d|\text{SHORT}(\text{Boris}, d)\} = \text{MAX}\{d|\text{SHORT}(\text{Doris}, d)\}$.

d. $\exists d\ \text{MAX}\{d|\text{SHORT}(\text{Boris}, d)\} = \text{MAX}\{d|\text{SHORT}(\text{Doris}, d)\} > d_{\text{short}}$ for some contextually specified standard d_{short}.

Now the crucial observation is that TALL and SHORT denote onto the same scale, but in opposite directions. The result is that the maximal degree to which a person is tall is automatically the maximal degree to which the person is short[2]. As a consequence, (7-a) and (7-c) are equivalent. Notice that (7-b) and (7-d) are not equivalent since the contextual standards for the long and short scales may well differ.

The next step in the reasoning is that since (7-a) and (7-c) are equivalent, they enter into semantic competition. This means that in some way they compete for which is the most efficient way of expressing their message. Now (7-c) uses a marked term, contrary to (7-a), and since there is no other difference between them, one can say (7-c) is more marked overall and therefore dispreferred[3]. As a result, (7-c) is blocked as a reading of (6-d) since the same message could have been conveyed more efficiently.

As a consequence, (7-d) is the only remaining reading, which means that (6-d) is disambiguated and, in the absence of other factors, will always be interpreted evaluatively. Compare, however, with (6-c) which can be evalua-

[2]Here in the former case "maximal" is understood relative to the canonical ordering on TALL scale, and in the latter relative to the inverse ordering, since SHORT is the antonym of TALL.

[3]The reason for this is not made explicit in Rett (2008b) but is plausible given earlier accounts of how marked terms are more rare and might take more time to process.

tive or not evaluative. Consequently, we cannot deduce from (6-c) that Boris and Doris are tall, which suffices to classify it as not evaluative.

2.2 Evaluativity in the comparative

Comparatives with unmarked adjectives such as in (8-a) are generally agreed upon not to be evaluative. On the other hand, there is disagreement in the literature as to whether comparatives with marked adjectives, e.g. (8-b), are evaluative.

(8) a. Boris is taller than Doris.

b. Boris is shorter than Doris.

Clark (1969) writes that "'Pete is worse than John' unambiguously impl[ies] negative evaluations of Pete and John" (p.391). That is, marked comparatives are seen as evaluative. However, Rett (2008b) argues that upon closer scrutiny, comparatives are not evaluative.[4]

Indeed, that comparatives are not evaluative follows fairly seamlessly from the analysis presented before for equatives. Let us assume that (8-b) is ambiguous between the evaluative and non–evaluative reading in (9-a) and (9-b).

(9) a. $\textsc{max}\{d|\textsc{short}(\text{Boris}, d)\} > \textsc{max}\{d|\textsc{short}(\text{Doris}, d)\}$

[4]Except, of course, comparatives with extreme adjectives, which are always perceived as evaluative. For example, *Tim is more moronic than Pete* clearly implies a judgement about the intelligence or absence thereof of the individuals in question. For the sake of simplicity, I will exclude these extreme adjectives from our discussion.

b. $\textsc{max}\{d|\textsc{short}(\text{Boris}, d)\} > \textsc{max}\{d|\textsc{short}(\text{Doris}, d)\} > d_{\text{short}}$

c. $\textsc{max}\{d|\textsc{tall}(\text{Boris}, d)\} > \textsc{max}\{d|\textsc{tall}(\text{Doris}, d)\}$

d. $\textsc{max}\{d|\textsc{tall}(\text{Boris}, d)\} > \textsc{max}\{d|\textsc{tall}(\text{Doris}, d)\} > d_{\text{tall}}$

Now the non–evaluative reading (9-a) cannot enter into competition with the reading in (9-c), where the marked adjective is replaced by its unmarked counterpart. The problem is that they do not mean the same thing, and therefore they do not enter into semantic competition. Thus, none of the readings is blocked and as a result, the marked comparative is not evaluative.

2.3 Critique of non–blocking analysis

The analysis presented in section 2.2 is appealing since the ambiguity that is ascribed to comparatives and evaluatives can explain why there are contexts in which they are evaluative and others in which they are not. Furthermore, this account is supported by the variability in the presuppositionality observed by Higgins (1977), who remarks that "comparatives containing marked adjectives from a ratio scale *can* be interpreted neutrally"[5].

However, the same studies' finding that marked comparatives are in general more presuppositional is not in line with the analysis. If we are to interpret this lack of experimental confirmation to problems in its design, then we will arguably also lose its support for Rett (2008b)'s analysis of comparatives.

[5]Emphasis added. The author defines ratio adjectives as those that can combine with a measure phrase and that have a clear zero point.

Also, the argument for the non–evaluativity of marked comparatives feels somewhat unsatisfying. The crucial step was to compare the reading (9-a) with (9-c). But the latter seems a rather surprising choice as competitor for (9-a). What we essentially have done is taken (10-a) and compared it with (10-b), concluding that they are not synonymous. On what grounds was *taller* even considered as a candidate? Notice that in general a sentence with *smaller* implies the *negation* of the same sentence with *larger*, so it seemed we could not have chosen a worse candidate for equivalence. And what is more, why is the synonymous (10-c) excluded as a candidate?

(10) a. Boris is shorter than Doris (non–evaluative)

b. Boris is taller than Doris (non–evaluative)

c. Doris is shorter than Doris (non–evaluative)

d. Boris is not taller than Doris (non–evaluative)

Rett (2008a) observes that apparently the switching of the arguments has blocked the semantic competition. Interestingly, a similar result might be derived from the *principle of the primacy of functional relations*(Clark, 1969). Or, perhaps a less strong restriction could be that pairs can enter in semantic competition only if they differ minimally, where *minimal difference* could be defined as a relation between sentences α and β that hold if (i) $\alpha \neq \beta$, and (ii) there is no sentence γ that is less different from α than β is[6] and that occurs at some point in a stepwise transformation from α to β.

[6]Of course some distance metric is implicit here. It could be a sort of Levenshtein distance on strings of words.

3 Experimental investigation

I will argue here that the proposed analysis of evaluativity needs to be founded on a more firm experimental investigation, so that our theories are informed not only by the intuition of those who design them, but also by more objective data revealing how people use the sentences in question.

3.1 Comparing comparatives and equatives: a first experimental proposal

For example, to the best of my knowledge, a presuppositional analysis such that of Higgins (1977) has not been performed for equatives. Higgins investigated various types of comparatives to see how much presupposition they carried relative to each other. In order to test the theory that has been presented here it will be crucial to gain insight into how presuppositional equatives are relative to comparatives. Rett (2008b) predicts that they are much stronger in what they presuppose. This can be tested by a paradigm adapted from Higgins (1977).

We present subjects an *acceptability* task. We make a list of pairs of non–extreme adjectives, one of which is marked and the other one not. For both adjectives in the pair we find two objects who clearly do not possess the denoted property[7]. For example, for the *tall–short* pair, we could take *dwarf*, *miniature* as candidates for (not) *tall* and *skyscraper*, *poplar* for (not) *short*.

[7]To ensure comparability with the Higgins (1977) study, one can copy the examples used.

We present subjects with sentences of the form "X is as A as Y," where A is an adjective and X and Y the candidates that clearly do not have property A. Subjects are then asked to rate the acceptability by clicking with a mouse somewhere on a bar ranging from 0 for totally unacceptable to 1 for totally acceptable.

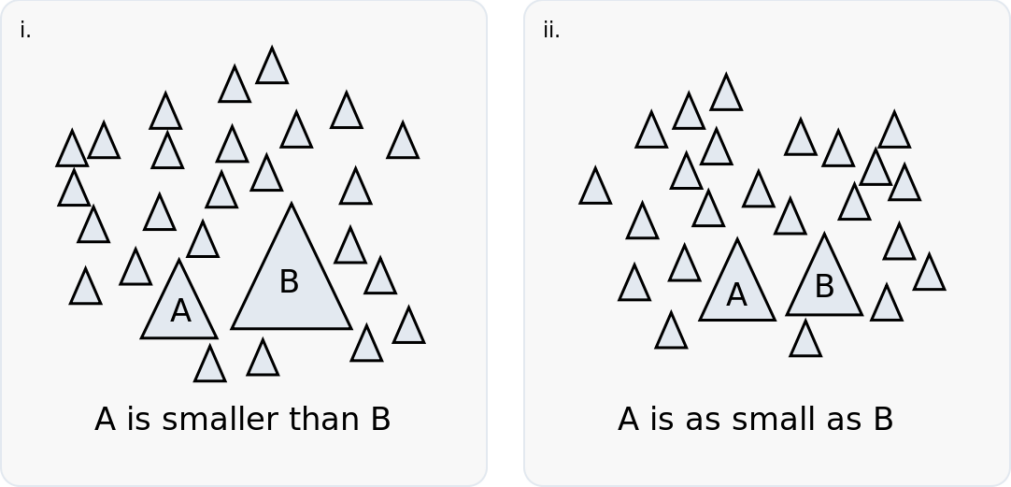
In addition to these we test the subjects on the marked–unmarked comparative from Higgins (1977)'s original study in order to ensure we replicate the effect and in order to provide a benchmark for the effect size of the equative.

Our theory predicts that the difference in acceptability between this equative marked–unmarked pair will be greater than that between the comparative marked–unmarked.

3.2 Context–sensitivity of comparatives and equatives

The problem in a Higgins (1977)–like approach to presuppositionality in comparatives and equatives is that we rely on subject's judgements independent of any context. This means that it is possible that the task becomes metalinguistic and therefore sensitive to many factors that come into play when people are asked to freely reflect on their opinion. For example, people might try to come up with a context or natural communication setting in which certain readings are appropriate, and thus their response would be a measure of their creativity much more than anything else. It would be preferable to address the issue or presuppositionality in a more direct way by making up

Figure 1: Equative and comparative embedded in context

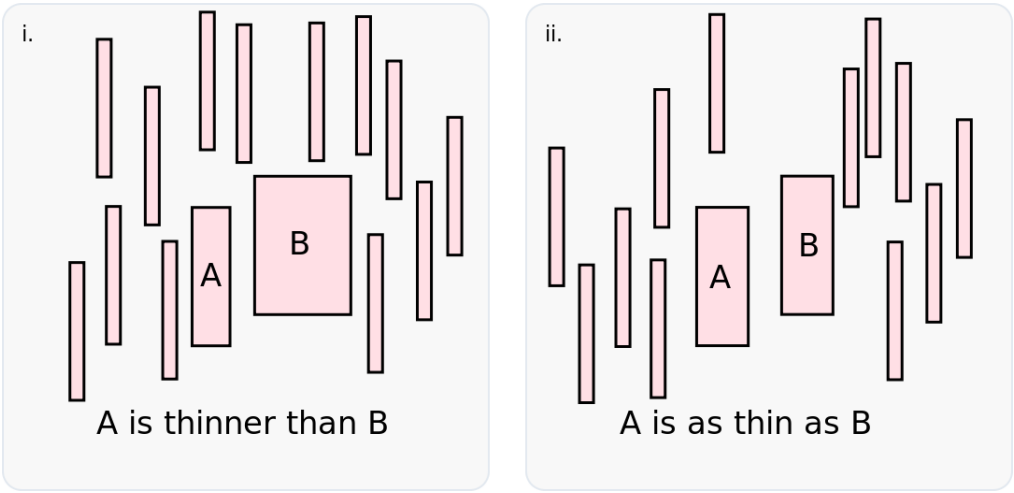


a concrete situation in which the judgements of people can be compared.

I propose an experiment in which a context is provided for two objects A and B that are compared for size by placing them in a field of smaller items. This means they are both relatively large. If our theory is correct, then that means that the equative *A is as small as B* will be dispreferred as a description when they are equal in size, since both are not small. However, when they differ in size, then *A is smaller than B* should be fine, since we can interpret it non–evaluatively and in that case it will be true. This is illustrated in figure 1 where the reader is invited to introspectively verify his own acceptability judgements.

A first part of this experimental program would be a pilot study where these pictures are given to subjects who are asked to rate them on a continuous scale. We predict that this will yield the same result as the acceptability

Figure 2: Using different adjective pairs to test the same predictions (or perhaps yield a different intuition?)



judgement task from the previous section, there the equative is significantly less acceptable than the comparative.[8] In order to make the purpose of the task less obvious to the participant, it will be sensible to include also the same cases but with a context of large objects. This will furthermore provide a baseline response against which the acceptability judgements of the two crucial cases can be compared. Also, the experiment can be peppered with other adjectives for which similar comparative and equative pictures can be drawn, for instance as shown in figure 2.

[8] I verified this informally with a naive subject who told me he hesitated tremendously to call the equative correct in the case of equating large objects in a small context by using *as small as*.

3.3 Picture–production paradigm

Once the results from this pilot study are established, we can move on to a more complex task in which we will simulate production by allowing the participant to choose from different utterance options which one best describes the picture in question.

In figure 3 the stimuli for the experiment are shown. Let us first consider the case of the equatives. We expect that in a LARGE context, both *A is as small as B* and *A is as large as B* are possible descriptions, since the latter can be interpreted non–evaluatively. In a SMALL context, *A is as small as B* is predicted to be not possible as a description since it can only be interpreted evaluatively, and *A* and *B* are not small, but large. This should be reflected in the overall participant's choice pattern.

Now in the case of the comparatives there are two possible answer schemas. Take the example of *A* being smaller than *B*. One schema (the *smaller* schema, cf. figure 3) proposes a choice between *A is smaller than B* and *A is larger than B*. These are the two sentences that are candidates for semantic competition in Rett (2008b). Notice that the latter is false; therefore all participants should choose the former if they are performing the task correctly.

In a second answer schema, referred to as *invert*, however, the participant can choose between *A is smaller than B* and *B is larger than A*. In this case, both answers are true in their logical sense. Rett (2008b) suggests that neither is presuppositional, and therefore neither is excluded for that reason.

This means that we expect to see no difference in choice pattern between these phrases in the LARGE context, nor in the SMALL context. If, however, the switching of the arguments is not as fundamentally disruptive as has been assumed, then we expect a preference for the use of the unmarked term in both contexts since apart from markedness of the term and the order of the arguments the utterances are identical.[9] Furthermore, reaction times might provide a clue as to the perceived difficulty or hesitation of the participants.

3.4 Time–course analysis of semantic competition

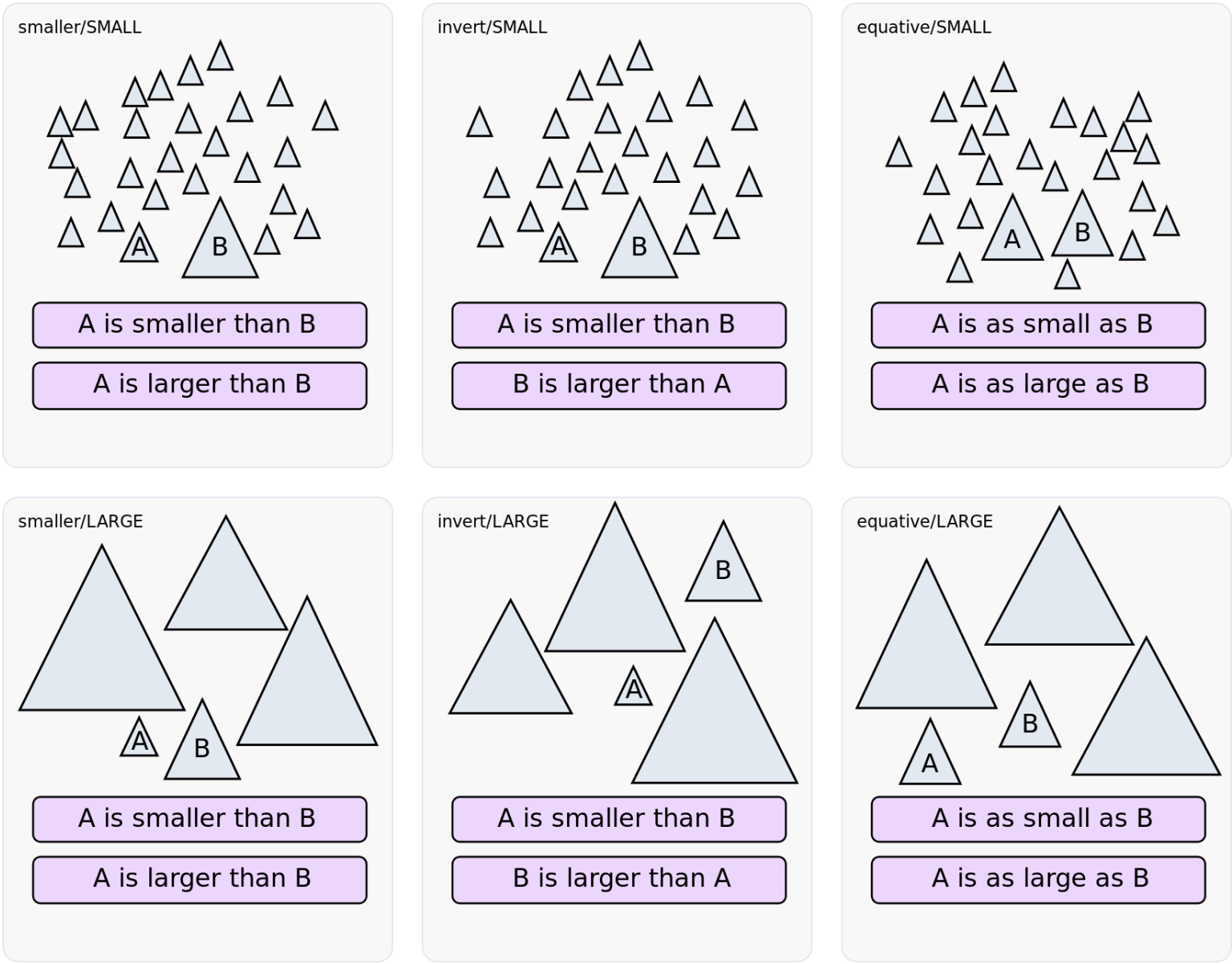
The semantic competition account provides a further possibility for experimental verification. The competition is in an abstract way comparable to the way Gricean implicatures are computed by a listener. Such implicatures are calculated as follows. If a listener hears a sentence ϕ and then considers a logically stronger sentence ψ that would have taken the same effort to produce, then he or she will conclude that the speaker thinks ψ is false. For otherwise, the speaker would have uttered ψ to be maximally informative.

If we assume for a moment that the speaker is intending to say that two objects A and B are equal in vertical size. Then he or she considers uttering one of (11). That is, the two are in competition. Now suppose that there is a Gricean–like maxim that dictates: *say what you have to say as efficiently as possible*, briefly: *be efficient*[10]. Now since (11) mean the same thing and

[9]The appeal of this experiment lies precisely in the comparison between the contexts in this case to be highly informative with respect to our theories.

[10]Perhaps this can be seen as a special case of the *maxim of manner* that requires us

Figure 3: Experimental design for the picture–production task



therefore convey exactly the same information, the usage of *short* is less efficient than *tall* since it is more marked. This means that the speaker will utter (11).

(11) a. A is as tall as B.

b. A is as short as B.

At this point, one should remark that nothing in the theory of semantic competition has committed us to this view that the competition unfolds in real time while the subject is preparing the utterance. This is analogous to how the theory of Gricean pragmatics does not imply that this implicature is calculated every time by the subject. For all we know it could also be hard–wired into the meaning of the word.

However, in the case of pragmatic implicatures Bott & Noveck (2004) showed that subjects who were told that *some* means "some or possibly all", i.e. the logical meaning of *some*, responded faster to verification studies than a different group of subjects who were instructed that it meant "some but not all", i.e. the pragmatic meaning. Also, subjects who were not instructed any particular meaning for *some*, responded according to the logical meaning more often when they were put under time pressure to respond. The authors conclude that calculating the pragmatic implicature takes time and that it is derived "on–line" every time the word *some* is used.

to be as clear as possible.

3.5 An experimental proposal for competition annihilation

This means that it is possible, though by no means necessary, that the semantic competition happens in real time. In this case we would be able to make people use the marked equative non–evaluatively.

The data from the experiment described in section 3.3 is needed for our first step. We investigate at what latencies subjects respond. Now a STRICT time limit is decided so that exactly 50% of the responses of the pilot subjects fall before and the rest after this time limit. Further, a LONG time limit is decided so that 90% of the responses is included.[11] Now the actual test subjects are divided into two groups. One group is given the STRICT time limit, the other the LONG time limit.

Our hypothesis that the semantic competition happens in a separate stage, after other picture–encoding decisions are taken, and therefore takes time makes the following prediction. Under the STRICT time limit, the equative in the SMALL context will be equally equally often described with *smaller* or *larger*, even though the pilot test presumably shows that it is dispreferred to use *smaller* in that context. However, in the LONG time limit, there should be a significant preference for *larger*, i.e. a replication of the results in the previous study without time–limit.

[11] We on purpose do not include all responses since (a) obviously there will be outliers, but also (b) it is important that subjects have at least some sense of time pressure in both cases, though in one case it is much more severe.

The same comparison can be made for the comparative in the *invert* condition (cf. figure 3). Depending on what effect we found in the earlier study without time pressure, seeing whether this *invert* condition is affected in the same way as the equative by increased time pressure will allow us to gain insight into the extent to which their evaluativity or not is comparable. Finally, the *smaller* condition (cf. figure 3) serves as a crucial control condition, since one of the examples is strictly wrong. This is vital if we would find that subjects choose equally often either response in the *invert* condition, which could be interpreted as a result of too high time pressure. Only when they do *not* respond at chance level in the *smaller* condition can we rule out this interpretation.

4 Conclusion

Certain degree scales are denoted into by pairs of opposite adjectives that are asymmetric in that one is the default, unmarked case and the other is its marked alternative. Phrases that relate two objects along a particular domain using such adjectives are often felt to be evaluative in the marked equative construction but not in the marked comparative, nor in any of the constructions using unmarked adjectives. In this paper, several experiments are proposed in full detail that can further clarify how these evaluativity patterns are used by human subjects, so that our finest semantic theories can be informed by rigorous empirical results.

References

Bott, Lewis, & Noveck, Ira. 2004. Some utterances are underinformative: The onset and time course of scalar inferences. *journal of memory and language*, **51**, 437–457.

Carpenter, Patricia A., & Just, Marcel Adam. 1975. Sentence comprehension: A psycholinguistic processing model of verification. *Psychological review*, **82**(1), 45–73.

Chase, W.G., & Clark, H.H. 1971. Semantics in the perception of verticality. *British journal of psychology*, **62**, 211–216.

Chase, W.G., & Clark, H.H. 1972. On the process of comparing sentences against pictures. *Cognitive psychology*, **3**, 472–517.

Clark, Eve V. 1972. On the child's acquisition of antonyms in two semantic fields. *Journal of verbal learning and verbal behavior*, **11**(6), 750 – 758.

Clark, Herbert H. 1969. Linguistic processes in deductive reasoning. *Psychological review*, **76**(4), 387–404.

Higgins, E. Tory. 1977. The varying presuppositional nature of comparatives. *Journal of psycholinguistic research*, **6**(3), 203–222.

Proctor, Robert W, & Cho, Yang Seok. 2006. Polarity correspondence: A general principle for performance of speeded binary classification tasks. *Psychological bulletin*, **132**(3), 416–442.

Rett, Jessica. 2008a. Antonymy and evaluativity. *In:* Gibson, M., & Friedman, T. (eds), *Proceedings of salt xvii*. CLC Publications.

Rett, Jessica. 2008b. *Degree modification in natural language*. Ph.D. thesis, Rutgers University.

Seymour, Philip H. K. 1974. Stroop interference with response, comparison, and encoding stages in a sentence-picture comparison task. *Memory & cognition*, **2**(1A), 19–26.