Shallowly accurate but deeply confused – how language models deal with antonyms[1]

Adèle HÉNOT-MORTIER — *Massachusetts Institute of Technology*

Abstract. Antonymic adjectives are subject to a variety of asymmetries regarding pragmatic inferences. The *Inference Towards the Antonym* (Horn, 1989; Krifka, 2007; Ruytenbeek et al., 2017; Gotzner et al., 2018) in particular, consists in deriving the antonym of an adjective *A* when encountering its negation (*not* A). Within a given antonymic pair, this inference is supposed to apply to a greater extent to negated *positive* adjective, as opposed to negated *negative* adjectives. This is especially true when the latter is morphologically transparent. In this paper, we test if recent Large Language Models capture this contrast using different probing methods. We conclude that some but not all models exhibit a contrast between positive and negative adjectives regarding the target inference, although (i) the observed contrasts are not readily interpretable at the level of word processing (ii) part of it may be explained by frequency differences (iii) more general expectations about the models' behavior regarding antonymic adjectives (parsing, reversing effect of negation) are not met. This casts doubt on the ability of such models to abstractly encode the concept of antonymy.

Keywords: antonymic adjectives, polarity, pragmatic inferences, language models, surprisal.

1. Background on adjective polarity

1.1. Semantics and pragmatics of antonymic adjectives

Antonymic adjectives, like (*tall*, *short*), (*nice*, *mean*), (*lucky*, *unlucky*) are roughly understood as semantic opposites. It has been observed that intuitively positive vs. negative adjectives pattern differently in several respects. First, only negative adjectives (abbreviated A^-) give rise to *Evaluativity Inferences* when used in equative and comparative constructions, as well as in questions (Bierwisch, 1989; Rett, 2015). This is shown in (1) and (2).

(1) a. John is as tall_{A^+} as Paul. $\rightsquigarrow$ Both may be tall or short.
b. John is as short_{A^-} as Paul. $\rightsquigarrow$ Both are judged to be short by the speaker.

(2) a. How tall_{A^+} is John? $\rightsquigarrow$ John may be tall or short.
b. How short_{A^-} is John? $\rightsquigarrow$ John is judged to be short by the speaker.

Second, negative (rather than positive) adjectives may feature overt negative morphology (Horn, 1989). The examples in (3) below illustrate this point.

(3) a. in-competent; im-modest; un-lucky; dis-honest …
b. *un-small; *im-messy; *un-poor; *dis-arrogant …

[1]I would like to thank Forrest Davis and the students of Fall 2022 Special Seminar on Computational Linguistics for their comments on earlier versions of this project. I also would like to thank the audience and reviewers of Sinn und Bedeutung 28, as well as the attendees of the poster session and reviewers of the 29th Architectures and Mechanisms of Language Processing conference. All errors are my own.

 In: Baumann, Geraldine, Daniel Gutzmann, Jonas Koopman, Kristina Liefke, Agata Renans, and Tatjana Scheffler (eds.) 2024. Proceedings of Sinn und Bedeutung 28. Bochum: Ruhr-University Bochum, 1079-1097.

Third, antonymic adjectives seem to differ in the inferences they lead to when placed under negation. Specifically, (4) shows that it appears easier to infer the antonym A^- of a negated positive adjective (abbreviated *not* A^+), than to infer the antonym A^+ of a negated negative adjective (*not* A^-) (Horn, 1989; Krifka, 2007; Ruytenbeek et al., 2017; Gotzner et al., 2018).

(4) a. He is not tall_{A^+}. $\rightsquigarrow$ He is fairly short_{A^-}.
b. He is not short_{A^-}. $\not\rightsquigarrow$ He is fairly tall_{A^+}.

The inference in (4a) was dubbed *Inference Towards the Antonym* (henceforth ITA); it will be the focus of this paper. An account of the ITA, due to Krifka (2007), is based on the idea that any two antonyms A and A' are pure logical opposites of each other, which means that by default (*not* A) $\equiv$ A' and (*not* A') $\equiv$ A. This implies (*not* A) $\models$ A' and (*not* A') $\models$ A, i.e. the ITA is a (logical) primitive. It can however be *mitigated* if A and A' vary in complexity. More precisely, if *not* A appears more complex than A', then there are good reasons to think that the speaker wanted to convey a meaning different from A' when uttering *not* A, i.e. (*not* A) $\not\models$ A'. This is summarized in (5), where CPLX refers to a measure of formal complexity.

(5) ITA Pragmatic Mitigation Condition (Krifka, 2007)
(*not* A) $\not\models$ A', if CPLX(*not* A) $\gg$ CPLX(A') $\diamondsuit$

This allows to explain how a contrast in ITA *can* arise, but does not yet predict *in which direction* it arises. Building on the additional assumption due to Büring (2007); Büring (2007) that all negative adjectives involve either overt or covert negation, Krifka derives the two equations in (6). Small caps NOT refers to morphological (and potentially covert) negation.

(6) Negative Adjectives Complexity Hypothesis (Büring, 2007; Büring, 2007)
$\forall\, A^-.\ A^- = \text{NOT-}A^+$, therefore:

$$\text{CPLX}(A^-) = \text{CPLX}(\text{NOT-}A^+) \quad \sim \text{CPLX}(\textit{not}\ A^+) \qquad (\spadesuit)$$

$$\text{CPLX}(\textit{not}\ A^-) = \text{CPLX}(\textit{not}\ \text{NOT-}A^+) \gg \text{CPLX}(A^+) \qquad (\clubsuit)$$

($\spadesuit$) states that *not* A^+ and A^- have the same degree of complexity. No mitigation should therefore occur for that pair, and the ITA should arise. In other words, *not* A^+ is expected to entail A^-. ($\clubsuit$) on the other hand, states that *not* A^- is significantly more complex than A^+. Pragmatic mitigation should therefore arise, leading *not* A^- and A^+ to have different meanings.

1.2. Previous experimental investigation of the ITA

In the study conducted by Ruytenbeek et al. (2017), the inference pattern presented in (4) was assessed in English and French using minimal pairs like (7). In such pairs, the presupposition trigger *too* is expected to lead to the inference that *not* $A^\pm$ in the first sentence and $A^\mp$ in the second sentence have similar meanings. Since this inference is licensed from *not* A^+ to A^-, but not so much from *not* A^- to A^+, (7a) is expected to be more felicitous than (7b).

(7) a. John is not tall_{A^+}. Paul is short_{A^-} too. (*not* A^+) $\rightsquigarrow$ A^-
b. # John is not short_{A^-}. Paul is tall_{A^+} too. (*not* A^-) $\not\rightsquigarrow$ A^+

In addition to testing the experimental validity of the basic contrast, and how it would correlate with independent measures of adjective polarity, Ruytenbeek et al. (2017) compared morphologically opaque pairs (e.g. *tall*/*short*) to morphologically transparent ones (e.g. *lucky*/*unlucky*). The goal was to investigate a refinement of the previous theory, based on the hypothesis that morphologically transparent pairs should lead to a stronger ITA contrast than morphologically opaque pairs. This stems from the idea that the decomposition A^-= NOT-A^+ is more salient when a negative adjective is transparent as opposed to when it is not–which in turn means that (♣) should hold even more unambiguously for morphologically transparent pairs. Therefore, a stronger contrast (signaled with a double hashmark) is expected in pairs like (8) as opposed to pairs like (7) above.

(8) a. John is not lucky$_{A^+}$. Paul is unlucky$_{A^-}$ too. (*not* A^+) $\rightsquigarrow$ A^-
b. ## John is not unlucky$_{A^-}$. Paul is lucky$_{A^+}$ too. (*not* A^-) $\not\rightsquigarrow$ A^+

Building directly on Ruytenbeek et al.'s study on human participants,we propose to test if some recent Large Language Models (henceforth LLMs) verify the two hypotheses laid out in (9). This is to our knowledge the first study of the ITA in the context of LLMs (though see Aina et al., 2019 for a study on negated antonymic adjectives on earlier models, and Cong, 2022 for a study on evaluativity and LLMs).

(9) **H1**: it should be easier to infer A^- from *not* A^+ than A^+ from *not* A^-.
H2: the contrast in ITA strength between (*not* A^+)/A^- and (*not* A^-)/A^+) is bigger with transparent ("T") pairs of adjectives as opposed to opaque ("O") ones.

2. Technical and methodological background

2.1. The Transformer architecture

Probabilistic models of language, being for the most part based on the Distributional Hypothesis (Harris, 1954), have been previously shown to display poor performances in rendering the meaning of antonymic adjectives, in particular w.r.t. their interaction with negation (Aina et al., 2019). Recent LLMs, which are based on the Transformer architecture (Vaswani et al., 2017) and in particular the concept of *attention*, supposedly allow for more complex contextual dependencies between words, and as such might better grasp the meaning of antonymic adjectives, and the functional behavior of negation. Such models are also based on a process called *tokenization*, which allows to break certain words into pieces (*tokens*). In the following we provide an overview of tokenization and multi-head self-attention.

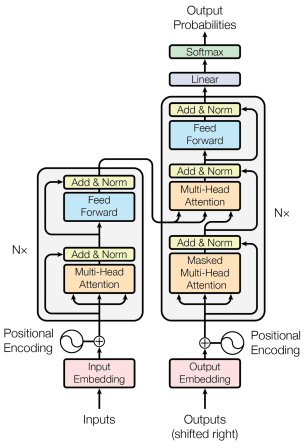
2.1.1. The tokenization process

Transformers operate on tokenized sentences, meaning, sentences whose words have been converted into one or several integers (*tokens*). Although it is not part of LLMs *per se*, tokenization remains crucial as it provides the models with interpretable inputs. The tokenization procedure relies on Byte-Pair Encoding (BPE), a process that creates tokens bottom-up from the set of

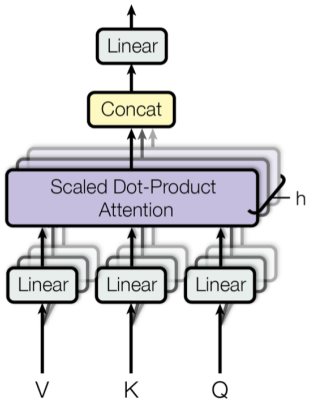
characters (unigrams) appearing in the training corpus (initial workspace), by iteratively (i) merging the most frequent bigram, (ii) putting this bigram back as a unigram (with its corresponding frequency) in the workspace. The process stops once a specific vocabulary size has been reached. Since BPE is based on *n*-gram frequencies, it is expected to capture a certain number of morphological regularities. For instance, the existence of the negative morpheme *un-* in English probably makes the frequency of the corresponding bigram (*u*+*n*) comparatively high, making it likely to be categorized as a complete token. We will see in Section 4.2 that this kind of prediction is at least partly borne out for the adjectives involved in our dataset. If BPE is "productive' in the sense that any new word can be tokenized using its output vocabulary, supplemented by the initial set of characters, and an extra "unknown" token for characters that did not belong to the initial set, this also entails that not all tokenizations will fully correspond to sensible morphological decompositions, either because some relevant morphemes are not identified, or because they are mistakenly identified in unexpected positions.

2.1.2. Multi-head self-attention

The main innovation of Transformers is the use of attention mechanisms, more specifically multi-head self-attention, as a core component of the network. Self-attention is a process that maps the representation of a given token t_j to an optimized mixture of the representations of the n surrounding tokens $\{t_i\}_{i\in[1;n]}$; the desideratum being that the weights of the mixture reflect how "relevant" those tokens are to t_j. *Multi-head* self-attention runs several such mechanisms ("heads") in parallel, allowing to capture different kinds of dependencies between tokens.



(a) The Transformer Encoder-Decoder architecture. Note that BERT only uses the encoder part, while GPT-2 only uses the decoder part.



(b) Detail of the multi-head self-attention architecture.

Figure 1: The Transformer architecture (taken from Vaswani et al., 2017).

Each head works as follows. First, the tokens $\{t_i\}_{i\in[1;n]}$ of the sentence are transformed ("embedded") into vectors $\{v_i\}_{i\in[1;n]}$ of dimension d_e. The goal of the self-attention head is then to map $\{v_i\}_{i\in[1;n]}$ to another set of vectors $\{y_i\}_{i\in[1;n]}$ containing more contextual information about

each other. This mapping relies on three main sets of parameters, packaged into three matrices whose weights are subject to optimization and vary for each different self-attention head: the Query matrix $\mathscr{Q}_{(d_k \times d_e)}$, the Key matrix $\mathscr{K}_{(d_k \times d_e)}$ and the Value matrix $\mathscr{V}_{(d_v \times d_e)}$. Focusing on one input token-vector v_j and its target contextual representation y_j, v_j is first transformed into a d_k-dimensional query vector q_j using $\mathscr{Q}$. n d_k-dimensional keys are obtained by multiplying each of the $\{v_i\}_{i\in[1;n]}$ by $\mathscr{K}$. A dot product is then performed between the query q_j and each of the keys to obtain a list of scalar numbers that are subsequently normalized to yield the weights $\{w_{ji}\}_{i\in[1;n]}$. Finally, n d_v-dimensional "values" are obtained by multiplying each of the $\{v_i\}_{i\in[1;n]}$ by $\mathscr{V}$, and those values get linearly combined together using the weights $\{w_{ji}\}_{i\in[1;n]}$. This mixture of values is itself a d_v-dimensional vector, namely y_j, the target contextual representation of t_j. This whole series of operations can be performed for all $j \in [1;n]$, and for each attention head $\{h_l\}_{l\in[1;m]}$, which leads to the more compact set of equations below. Note that the matrices $\mathscr{Q}$, $\mathscr{K}$, $\mathscr{V}$ now covary with the attention head h_l to model different kinds of contextual dependencies, and that the outputs of the m heads are combined and weighted by a matrix $\mathbf{W}$. Moreover, as Figure 1 shows, several (N) multi-head self-attention modules actually get *stacked* in the global architecture.

$$\mathbf{E} = \left[\left[\begin{array}{c} \\ v_1 \\ \\ \end{array}\right] \cdots \left[\begin{array}{c} \\ v_N \\ \\ \end{array}\right]\right] \qquad \begin{array}{rcccl} \mathbf{Q}_l & = & (\mathscr{Q}_l\mathbf{E})^T & : & N \times d_k \\ \mathbf{K}_l & = & \mathscr{K}_l\mathbf{E} & : & d_k \times N \\ \mathbf{V}_l & = & (\mathscr{V}_l\mathbf{E})^T & : & N \times d_v \end{array}$$

$$\forall l \in [1;m].\ h_l = \text{softmax}\left(\frac{\mathbf{Q}_l\mathbf{K}_l}{\sqrt{d_k}}\right)\mathbf{V}_l$$

$$\textsc{MultiHead}(\mathbf{Q},\mathbf{K},\mathbf{V}) = [h_1 \ldots\ h_m]\mathbf{W}$$

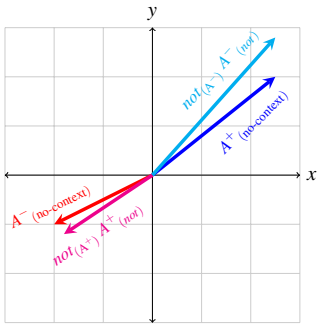


Figure 2: Idealized representation of the word-vectors of an antonymic pair and of their respective negations.

In practice, the output of vanilla LLMs is generative, which means that LLMs predict tokens given a certain context, by assigning them probabilities. Left-to-right models (e.g. the GPT family, cf. Radford et al., 2018; Radford et al., 2019; Brown et al., 2020) compute token representations and probabilities in a left-to-right fashion, while bidirectional models (e.g. the BERT family, cf. Devlin et al., 2018; Liu et al., 2019) do so using both left and right contexts.

2.2. The challenge of negated antonymic adjectives

Antonymic adjectives, and more so negated antonymic adjectives, pose a double problem to statistical models of language. The first problem is that of *grounding* (cf. Bender and Koller, 2020 a.o.). Since LLMs are simply trained to predict tokens, it is notoriously hard for them to capture intuitions about properties of the physical world, such as weight or size (though see Grand et al., 2022 for a discussion on the achievement of earlier models on nominals). This of course is problematic for adjectives, since many of them are highly context-dependent, and elements of an antonymic pair will often appear in similar distributional environments (Charles and Miller, 1989; Justeson and Katz, 1991). The second issue comes from the effect of nega-

tion. In formal semantics, negation is typically seen as a function that takes a proposition or predicate as argument, and return its opposite (proposition with opposite truth conditions, or complement set). LLMs however, treat any word as a vector, and therefore there is no way to properly "apply" the representation of negation to that of its argument in order to "reverse" it. One way for negation to alter its argument is in fact attention: as outlined in the previous section, LLMs derive contextual representations of words within a given sentence, so, ideally, we might expect negation to modify the representation of the adjective (and vice versa, in bidirectional architectures) in such a way that the representation of the negated adjective (typically seen as the mean of the representations of its constitutive tokens) becomes more or less close to the representation of its antonym, depending on polarity. This is schematized in Figure 2, where indices represent the context (assumed bidirectional) used to derive the representation of each token. Note however that this idealization puts a very high pressure on the contextual aspect of representations: if $\overrightarrow{not\ _{(\mathrm{A}^{\pm})}\mathrm{A}^{\pm}_{(not)}} = 1/2\left(\overrightarrow{not_{(\mathrm{A}^{\pm})}} + \overrightarrow{\mathrm{A}^{\pm}_{(not)}}\right) \simeq \overrightarrow{\mathrm{A}^{\mp}}$ [2] then the contextual representation of *not* given any adjective of the antonymic pair is $\overrightarrow{not_{(\mathrm{A}^{\pm})}} \simeq 2\overrightarrow{\mathrm{A}^{\mp}} - \overrightarrow{\mathrm{A}^{\pm}_{(not)}}$ and the difference between the two contextual representations of *not* becomes proportional to the difference of the context-free representations of the two antonyms, which arguably is non-negligible: $\overrightarrow{not_{(\mathrm{A}^{+})}} - \overrightarrow{not_{(\mathrm{A}^{-})}} \simeq 3\left(\overrightarrow{\mathrm{A}^{-}} - \overrightarrow{\mathrm{A}^{+}}\right)$. We will see in Section 3.4 that this kind of constraint on contextual representations is not satisfied by the LLMs under study. The next two sections introduce two ways of probing the capacity of LLMs to successfully encode the semantics of adjectives.

2.3. Evaluating the linguistic performance of LLMs with surprisal

In human studies, the negative log-probability (surprisal) of a given word in a given context was shown to correlate with general processing effort (Hale, 2001; Levy, 2008). By extension, surprisal was taken as a reasonable proxy for syntactic acceptability when investigating the "linguistic" behavior of statistical models of language (Wilcox et al., 2018; Futrell et al., 2019; Wilcox et al., 2023) w.r.t. a variety of phenomena, among which filler-gap dependencies and island effects. The assumptions of this line of work are summarized in the equations in (10), where t_i denotes a token and $C(t_i)$ its context. For left-to-right models, C will denote the left context only, while for bidirectional models, C will denote both the left and right context. We will use the same kind of methodology in this paper except that we will assume that the measure of ACCEPTABILITY defined in (10) can, in the sentences at stake, reflect pragmatic acceptability.

(10)

$$\begin{aligned}
\textsc{Surprisal}(t_i, C(t_i)) &= -\log\left(\mathbb{P}(t_i|C(t_i))\right) \\
\textsc{Acceptability}(t_i, C(t_i)) &\simeq -\textsc{Surprisal}(t_i, C(t_i)) \\
\textsc{Acceptability}(t_1 \ldots t_N, C) &\simeq -\sum_{i=1}^{N} \textsc{Surprisal}(t_i, C(t_i))
\end{aligned}$$

[2] Given our hypothesis about the ITA, this last equality holds more for *not* A^{+} than for *not* A^{-}. This is illustrated in Figure 2: the vector of A^{-} is slightly closer to that of *not* A^{+} than the vector of A^{+} is to that of *not* A^{-}.

2.4. Evaluating LLMs on logical inferences

It is also possible to evaluate certain LLMs on logical inferences without appealing to measures of surprisal. In that case, the models are fine-tuned to perform at specific kind of classification task called *Natural Language Inference* (NLI). Fine-tuning consists in keeping most of the parameters of the model untouched, while adding (and training) an extra final layer suited for the particular task at stake. For NLI, the task typically consists in deciding if two sentences are in a relation of logical entailment, contradiction, or logically independent, by outputing a probability. Although it appears more direct than a surprisal-based assessment, this kind of task relies on the capacity of LLMs to transfer a "knowledge" acquired on the general instances of entailment encountered during training, to the particular case of the ITA.[3]

3. Experiments

3.1. Setup

The code used for the experiments is available here. First, a dataset comprising 107 pairs of English antonymic adjectives was manually created. There was some degree of redundancy in the adjectives used across pairs, due to synonymy. For instance, the positive adjective *kind* was paired to the opaque negative adjective *mean*, but also, to the transparent negative adjective *unkind*. The dataset contained a total of 48 transparent ("T") pairs, and 59 opaque ("O") pairs. The experiments involved three main tasks. Task 1 focuses on surprisal measures to assess differences in ITA strength. Task 2 probes NLI models to directly measure ITA strength *via* entailment probabilities. Task 3 compares the contextual vector representations assigned by LLMs to antonyms and their respective negations to determine if contrasts in ITA strength translate into model-internal topological regularities. In Tasks 1 and 3, four models (all in their "Large" version from Huggingface) were tested: GPT-2 (Radford et al., 2019), XLNet (Yang et al., 2019), BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019). As mentioned earlier, GPT-2 is purely left-to-right; while BERT and RoBERTa are bidirectional. Lastly, XLNet features a left-to-right architecture but its objective function allows it to incorporate some bidirectional dependencies during training, which, arguably, makes it combine the best of both worlds.[4] In the second task, two models fine-tuned on the MNLI dataset (Williams et al., 2018) were assessed: RoBERTa (supposedly better than BERT) and DeBERTa (supposedly better than RoBERTa, He et al., 2020). Both were in their "Large" version.

3.2. Task 1: comparing measures of surprisals in minimal pairs

This task aimed at testing surprisal contrasts at the word- and sentence-level, on minimal pairs following the template in (11), inspired by Ruytenbeek et al.'s original stimuli.[5] All pairs of

[3]This holds even more that one of the most popular NLI dataset used to fine-tune models, MNLI (Williams et al., 2018), contains very few instances of pure pragmatic inferences, as observed by Jeretic et al. (2020).

[4]Bidirectional models are expected to be overall better at modeling natural language, since not all kinds of dependencies are purely left-to-right. Those models however, are trained on a masked language modeling objective,

sentences were counterbalanced for gender (by swapping the pronouns) and "filled" with the 107 possible (A^+, A^-) antonymic pairs. For each minimal pair, we collected differences in sentence-level and word-level[6] surprisals using the Python `minicons` library (Misra, 2022).

(11) "Anteposed *too*" template

a. He is not A^+. She too is A^-.

b. # He is not A^-. She too is A^+.

(12) spells out how the hypotheses introduced in (9) translate in terms of total sentence surprisal contrasts for the template in (11).

(12) **H1:** SURPRISAL(11b) - SURPRISAL(11a) > 0
H2: $\text{SURPRISAL(11b)}_{A \in T}$ - $\text{SURPRISAL(11a)}_{A \in T}$ > $\text{SURPRISAL(11b)}_{A \in O}$ - $\text{SURPRISAL(11a)}_{A \in O}$

Figure 3a shows that all models but one (RoBERTa) exhibit a significant contrast in surprisal as a function of adjective polarity, with effect sizes varying from small to large.[7] This is in line with H1. Figure 3b shows that with GPT-2 and BERT, H1 is also individually verified by both the T- and O-group (with small to large effect sizes).[8] GPT-2 additionally appears to verify H2, meaning, the T-group is associated to a bigger contrast in ITA strength than the O-group, with a small effect size.[9] BERT only marginally verifies this prediction after corrections. This constitutes preliminary evidence that some LLMs capture the contrast in ITA strength *vis à vis* adjective polarity (H1) and its interaction with morphological transparency (H2). Remarkably, the two models that were supposedly more robust on general linguistic benchmarks, RoBERTa and XLNet, appear to perform less well than the basic models on this task.

---

which consists in replacing input tokens one at a time by a dummy token `MASK`, and learning to predict the original token using its bidirectional context. This causes this family of models to get worse at fine-tuning (which does not involve any artificial `MASK` token); and, also, this makes such models unable to capture joint probabilities in their prediction of the masked tokens, due to them being predicted only on the basis of their respective contexts.

[5]In addition to the template in (11), we tested a template in which *too* appeared after the second adjective, and a template without *too* but with the predicate *mean* coordinating the two propositions.

(i) "Postposed *too*" template

a. He is not A^+, and she is A^- too.

b. # He is not A^-, and she is A^+ too.

(ii) "Meta" template

a. He is not A^+ means that he is A^-.

b. # He is not A^- means that he is A^+.

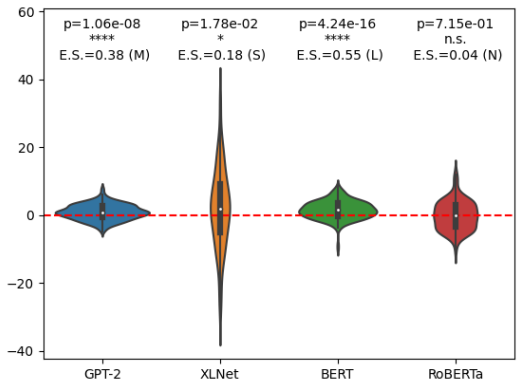
The first template is closer to Ruytenbeek et al.'s original stimuli but is doing less justice to the left-to-right models (for which processing the presupposition trigger *too* before the second adjective might be crucial). The second template was used to neutralize the role of presuppositions in the semantic judgment, as it is yet unclear whether LLMs reliably "compute" presuppositions in the first place (Jeretic et al., 2020). Results for those templates can be generated using the project notebook, and do not fundamentally differ from those presented here.

[6]If a word was segmented into several tokens, its surprisal was computed by simply summing the surprisals of its constituent tokens.

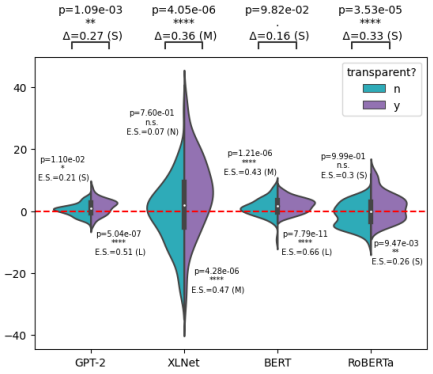
[7]This result is robust across templates for GPT-2 and BERT.

[8]Not 100% robust across templates: the O-group in the postposed *too* paradigm with GPT-2, and the T-group in the "meta" paradigm with BERT, failed to reach significance.

[9]Robust across all three templates–cf. footnote 3.2 for what the other templates were.



(a) Testing H1. p-values[10] computed using one-tailed, Holm-Bonferroni-corrected Wilcoxon tests, effect size E.S. = $\frac{|z|}{\sqrt{n}}$.



(b) Testing H2 (T- vs. O-group). Within-group p-values computed using one-tailed, HB-corrected Wilcoxon tests; between-group using HB-corrected Mann-Whitney U-tests, and Cliff's Δ as effect size.

Figure 3: Paired differences in total sentence surprisal between (11b) and (11a).

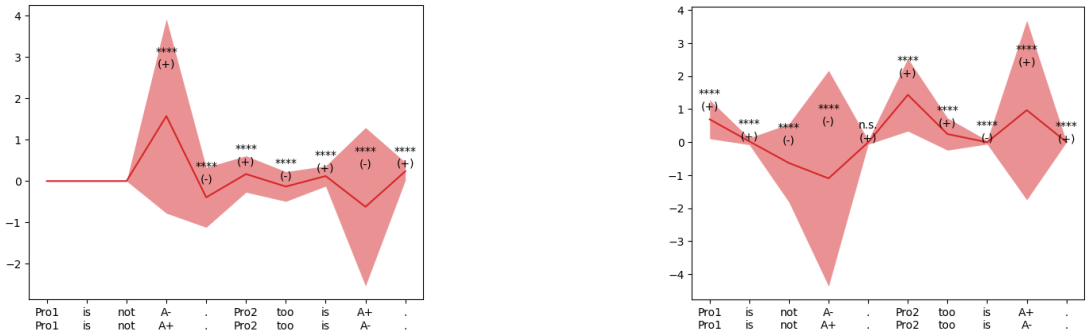
Let us now focus on what the two best-performing models do at the word-level. From a language processing standpoint, we expect the positive contrasts in surprisal hypothesized at the sentence-level in (11) to be mainly driven by the occurrence of the second adjective. This adjective is expected to be relatively unsurprising when *negative* (due to the comparatively stronger ITA triggered by the negated *positive* adjective in the preceding sentence) and more surprising when *positive* (due to a weaker or absent ITA in the preceding sentence). This is summarized in (13) below, where A_2 refers to the second adjective in (11a)/(11b). This prediction however, might be influenced by whether the model under study is left-to-right or bidirectional. Indeed, since in bidirectional models the conditional probability of each token is computed given its left and right context, the surprisal contrast expected on the second adjective might spread to other elements preceding it and "interacting" with it *via* attention, typically, the presupposition trigger *too*, the predicate *mean*, or the first adjective.

(13) **H1:** SURPRISAL(A_2, 11b) - SURPRISAL(A_2, 11a) > 0
H2: SURPRISAL(A_2, 11b)$_{A_2 \in T}$ - SURPRISAL(A_2, 11a)$_{A_2 \in T}$ >
SURPRISAL(A_2, 11b)$_{A_2 \in O}$ - SURPRISAL(A_2, 11a)$_{A_2 \in O}$

Figure 4a shows that GPT-2 treats A^- as significantly more surprising than A^+ in positive contexts (position 9, second adjective), but, even more so, in negative contexts (position 4, first adjective). The contrast in surprisal observed at the sentence-level for GPT-2 therefore seems to be driven by the first adjective, and not the second adjective, contrary to intuitions about the ITA, but perhaps consistent with a general avoidance for doubly negated (and therefore marked) structures, following Büring's Negative Adjectives Complexity Hypothesis. With BERT, Figure 4b shows that the pattern gets partly reversed: A^+ appears significantly more surprising than A^- in both positions, but, more remarkably perhaps, a surprisal contrast arises at the level of the subject of the second sentence (which remained the same in both sentences of a given

[10] p-value coding scheme: [.0001; −∞] ≡ ****; [.001; .0001] ≡ ***; [.01; .001] ≡ **; [.05; .01] ≡ *; [.1; .05] ≡ ·.

minimal pair!). This is consistent with the idea that bidirectional models tend to "spread" surprisal contrasts to neighboring "relevant" tokens; however the reason why the subject pronoun should be relevant to the adjective polarity contrasts remains quite obscure. Other relevant candidates, such as the presupposition trigger *too*, show a significant, although comparatively smaller, surprisal contrast. In sum, a word-level assessment of surprisal contrasts for the best-performing models suggests that the global effect witnessed at the sentence-level was driven by elements of the sentence which intiuitively were not predicted to be triggering the linguistic contrast. This in turn suggests that LLMs may rely on more superficial cues (such as bare frequencies, and perhaps, a derived concept of markedness) to assign sentences probabilities. Before digging even further into the LLMs' contextual representations of antonymic adjectives, we explore in the next section another method of assessing the strength of the ITA in minimal pairs.



(a) GPT-2 (b) BERT

Figure 4: Paired word-by-word differences in surprisal between (11b) and (11a), p-values computed using Wilcoxon tests. The red line tracks the mean surprisal for each word and the red enveloppe tracks the standard deviation.

3.3. Task 2: comparing entailment probabilities between minimal pairs

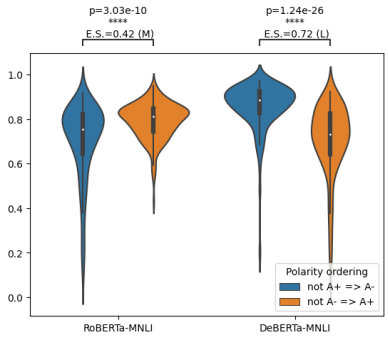
As outlined in Section 2.4 the contrast in (11) can be assessed using models fine-tuned to perform NLI. Those models are expected to associate the entailment patterns in (14) to a measure of probability reflecting how likely the relevant entailment is to hold for a particular pair of adjectives. (15) summarizes the specific predictions of (9) for this task, with $p^{\pm}_{A\in X}$ being the probability of entailment from *not* $A^{\pm}$ to $A^{\mp}$ when $A^{\pm}$ belongs to group X (T or O)

(14) a. He is not $A^+ \overset{p^+}{\Longrightarrow}$ He is A^-.
b. He is not $A^- \overset{p^-}{\Longrightarrow}$ He is A^+.

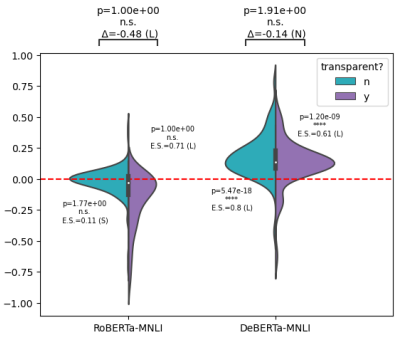
(15) **H1:** $p^+ - p^- > 0$
H2: $p^+_{A\in T} - p^-_{A\in T} > p^+_{A\in O} - p^-_{A\in O}$

Figure 5a shows that entailment scores are overall high for both models and both entailment schemes. Yet, only one of the two models (DeBERTa-MNLI) correctly predicts the inference in (14a) to be stronger than the one in (14b), in line with H1. The other model, RoBERTa-MNLI in fact predicts the opposite pattern. This negative result is consistent with the poor performance of the non-fine-tuned RoBERTa model in the previous task. Figure 5b shows that DeBERTa verifies H1 for the T- and O-groups individually (both with large effect sizes), but

also that there is no significant difference in entailment strength between the two groups, which means that H2 fails to be supported. Overall, this inference task is not extremely explanatory, because it does not allow to determine if the models are drawing the desired inference for the "right" reasons. The next section is an attempt to better delineate what the basic models do under the hood, by analyzing the contextual representations of antonymic adjectives and their negations in the models' vector spaces.



(a) Testing H1. Probabilities of entailment for (14a) vs. (14b) on two LLMs fined-tuned for NLI. Same tests are in previous tasks.



(b) Testing H2. Paired differences in entailment probabilities between (14a) and (14b), T- vs. O-group. Same tests as in previous tasks.

Figure 5: Differences in entailment probabilities between (14a) and (14b).

3.4. Task 3: comparing vector representation of adjectives and their negations

Recall Figure 2, which illustrated what one should expect of two-dimensional, linguistically sensible contextual vector representations of antonymic adjectives and their negations. This Figure showed that A^+ and *not* A^- on the one hand, and A^- and *not* A^+ on the other hand, should cluster together and that, additionally, A^- and *not* A^+ should be closer to each other than A^+ and *not* A^-, due to the expected differences in ITA strength. The most common measure of semantic proximity used in word embeddings is cosine similarity, defined below, which corresponds to the measure of the angle between two vectors.[11] If H1 and H2 translate into the LLMs' contextualized vector space, we then expect the inequalities in (16) to hold.

(16) $\textsc{CosSim}(\vec{v}_1, \vec{v}_2) = \frac{\vec{v}_1 . \vec{v}_2}{||\vec{v}_1|| \times ||\vec{v}_2||} \in [-1; 1]$

H1: $\textsc{CosSim}(\overrightarrow{\textit{not } A^+}, \overrightarrow{A^-}) - \textsc{CosSim}(\overrightarrow{\textit{not } A^-}, \overrightarrow{A^+}) > 0$

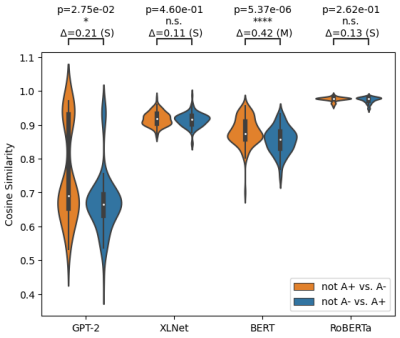
H2: $\textsc{CosSim}(\overrightarrow{\textit{not } A^{\pm}_T}, \overrightarrow{A^{\mp}_T}) - \textsc{CosSim}(\overrightarrow{\textit{not } A^{\mp}_T}, \overrightarrow{A^{\pm}_T}) >$
$\textsc{CosSim}(\overrightarrow{\textit{not } A^+_O}, \overrightarrow{A^-_O}) - \textsc{CosSim}(\overrightarrow{\textit{not } A^-_O}, \overrightarrow{A^+_O})$

For each model, we constructed vectors for $A^{\pm}$ and *not* $A^{\pm}$, by averaging the representations of the second-to-last layer of the model obtained for each token.[12] Figure 6a shows that all models

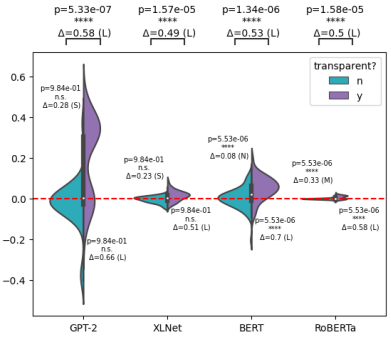
[11] Two vectors pointing in the same direction will have a cosine similarity of 1, regardless of their respective lengths, while two vectors pointing in opposite directions will have a cosine similarity of -1. Orthogonal vectors have a cosine similarity of 0.

[12] Because some models tokenize words differently depending on whether they are preceded by a white space or

associate adjectives and the negation of their antonym to fairly high cosine similarities. GPT-2 and BERT moreover treat *not* A^+ and A^- as closer to each other than *not* A^- and A^+, with small to medium effect sizes, in line with H1 and the results of Task 1. Figure 6b additionally shows that BERT individually verifies H1 for both the T- and O-groups, as well as H2.



(a) Absolute cosine similarities for both adjective orderings. *p*-values computed using one-tailed, Holm-Bonferroni-corrected Wilcoxon tests, effect size E.S. = $\frac{|z|}{\sqrt{n}}$.



(b) Paired differences in cosine similarities, T- vs. O-group. Same tests and corrections as in previous tasks.

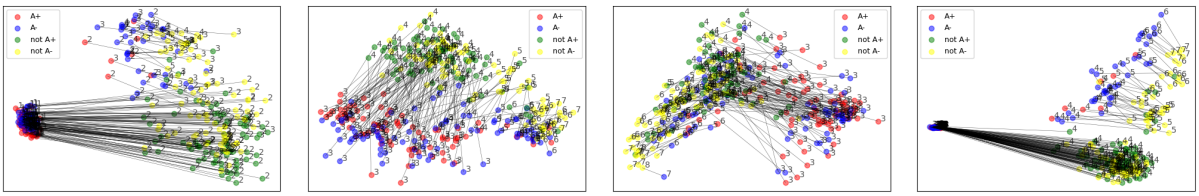
Figure 6: Differences in cosine similarities between $(\overrightarrow{\textit{not A}^+}, \overrightarrow{A^-})$ and $(\overrightarrow{\textit{not A}^-}, \overrightarrow{A^+})$.

These results are quite encouraging overall and suggest that the models which captured the desired surprisal contrasts in Task 1, encode antonymic adjectives and their negation in a somewhat sensible way, as well. This however, has to be nuanced with another fairly concerning aspect of the LLMs' contextual embeddings, visible in Figure 7 below, whereby bare antonyms (blue and red dots), and their negations (yellow and green dots) respectively end up clustering together in a 2D space where the dimensions that are retained are the ones that explain the most variance of the data. This clustering effect is evidently bigger than the one measured previously by comparing cosine similarities in higher dimensional spaces, and shows that the "reversing" effect of negation was not encoded by the models, thus replicating the negative result of Aina et al. (2019) for earlier models. Another concerning aspect of those 2-dimensional projections is the fact that the distributions of the vectors appear highly sensitive to the number of tokens they are derived from–this is particularly visible in the case of GPT-2 and RoBERTa for bare adjectives. This might also explain the bimodal aspect of the distribution of cosine similarities for GPT-2 in Figure 6, and calls for a more in-depth analysis of the LLMs' tokenization strategies.

As an interim summary, it appears that some, but not all of the LLMs under study captured the effect of adjective polarity on the Inference Towards the Antonym, and did so, at the level of sentences (*via* surprisal measures) and at the level of contextualized word representations (*embeddings*). The measuring of word-level surprisals, as well as a broader analysis of the LLMs' contextual embeddings, however cast doubt on whether LLMs "draw" the target inference for

not, we included an initial space before all the bare adjectives, to ensure they would be tokenized in the same way as they would be after negation. We also tried different vector extraction methods, in particular last-layer extraction (generally dispreferred due to the tendency of the last layer to encode information that is too task-specific) and summing of the last 4 layers (empirically better on certain benchmarks). Both methods led to comparable results as the one we retained in the main text, although the last-layer method led to slightly worse plots and *p*-values.

the right reason. In the next section, we explore two potential confounding factors: adjective frequencies and possible biases caused by the tokenization procedure.



(a) GPT-2 (b) XLNet (c) BERT (d) RoBERTa

Figure 7: Two-dimensional reductions (*via* Principal Component Analysis) of the contextualized representations of A^+, A^-, and their respective negations. The numbers indicate the total number of tokens (including start/end/separating tokens) each vector is derived from.

4. Analysis of confounding factors

4.1. Adjective frequency

Since the training of Transformer models relies on statistical regularities, one might wonder if the effects observed are not just artifacts of frequency differences between positive vs. negative adjectives, and/or transparent vs. opaque adjectives. Can adjective frequencies explain the behavior of the LLMs under study w.r.t. Tasks 1 (surprisal) and 3 (inference)? To answer this question, we used a dataset from Kaggle[13] gathering the frequencies of the 1/3 million most frequent English words on the Web. This dataset was derived from the Google Web Trillion Word Corpus, distributed by the Linguistic Data Consortium (Brants, Thorsten and Franz, Alex, 2006). Even if the composition of this dataset might differ from those of the datasets used to train the LLMs at stake,[14] we took it to be a sufficiently good approximation. This allowed us to extract the frequencies of all the adjectives from our dataset, which we further log-transformed and normalized.

Top 10 least frequent adjectives	Top 10 most frequent adjectives
ungraceful	just
uncommunicative	good
unambitious	well
unsocial	old
graceless	social
uncharitable	young
discourteous	popular
dextrous	fun
immodest	short
untruthful	bad

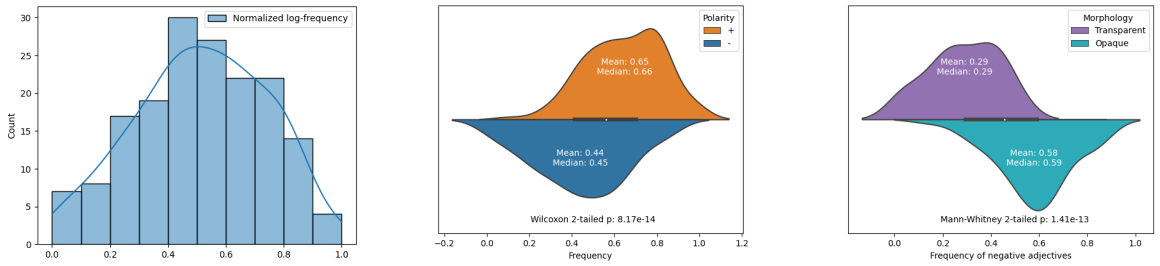
Table 1: 10 least and most frequent adjectives according to the Kaggle dataset.

Table 1 and Figures 8a-8c illustrate the distribution of those normalized frequencies. Figure 8b shows that positive adjectives are overall more frequent than negative ones, within each pair (2-tailed paired Wilcoxon p=8.17e-14) as well as globally (2-tailed Mann-Whitney p=2.08e-11). This might be partly explained by the fact that more positive adjectives from the dataset have homonyms, and as such got their frequencies increased, as opposed to negative ones, which

[13] Dataset available at https://www.kaggle.com/datasets/rtatman/english-word-frequency.

[14] As an example, GPT-2 was trained on BookCorpus, which comprises 7,000 self-published independent books, and a curated Web corpus called WebText involving 8 million web pages. BERT was trained on BookCorpus and Wikipedia.

in almost half of the cases featured negative morphology specific to adjectival forms. *Just* and *well* in the top 10 most frequent positive adjectives in Table 1 are examples of such ambiguous positive adjectives. Figure 8c shows that within the class of negative adjectives, transparent ones appear less frequent on average than opaque ones (2-tailed Mann-Whitney p=1.41e-13). This again, might be partly explained by the potential for homonymy of O-adjectives.



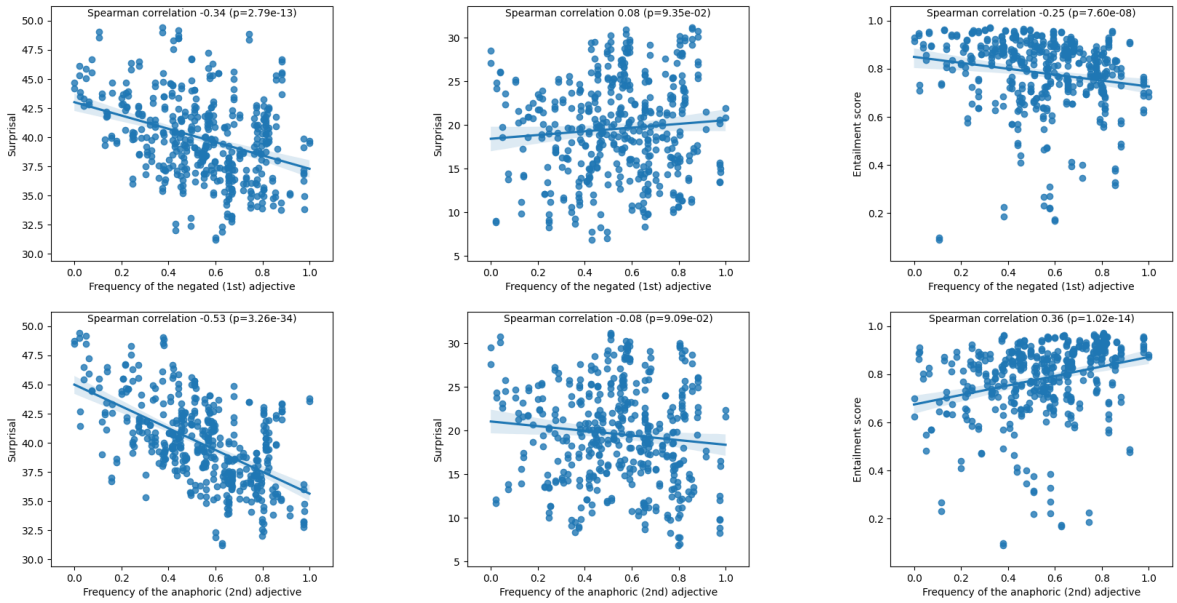
(a) All adjectives (b) By polarity (c) By transparency

Figure 8: Distribution of the normalized log-frequencies of the adjectives from our dataset.

Given these preliminary observations, we tried to assess the degree of correlation between total sentence surprisal measures (from Task 1) or entailment scores (from Task 2) on the one hand, and, on the other hand, the normalized log-frequencies of either the first (negated) adjective, or the second ("anaphoric") adjective in sentences like (11). To this end, we focused on the best-performing models for each Task. The intuitive expectation is that sentence surprisal measures should anti-correlate with the frequency of both adjectives, since surprisal convaries with the negative conditional probabilities of the tokens appearing in the sentence. Blocks (a) and (b) in Figure 9 show that this prediction is rather strongly verified for GPT-2, but not for BERT. This appears consistent with the word-by-word surprisal plots of Figure 4, which showed that the surprisal contrast with GPT-2 was driven by this model being overall more "surprised" at negative (i.e. less frequent) adjectives than positive (i.e. more frequent) ones, and that BERT weakly followed the opposite pattern. Regarding entailment scores, the prediction is less clear but we might expect more frequent adjectives in the conclusion to boost the probability of entailment. The lower plot of the (c) block in Figure 9 shows that this intuition is verified: when the adjective present in the conclusion becomes more frequent, the entailment score tends to increase, as well. It also seems that more frequent adjectives in the *premise* tend to make the entailment scores decrease (upper plot of the (c) block). This analysis suggests that GPT-2 and DeBERTa may heavily rely on bare adjective frequencies to produce the desired contrasts in surprisal and entailment probabilities, respectively.

4.2. Tokenization and morphology

A last aspect of LLMs that may require further investigation is their tokenization procedure. As briefly outlined in Section 2.1, the input of Transformer models is a tokenized string, whose tokens may or may not coincide with actual morphemes. Tokenizers vary across models. Are the tokenized inputs formed out of our adjective dataset any close to morphologically-segmented data? Does the number of tokens of positive vs. negative adjectives reflect differences in formal complexity that can in turn influence surprisal or inference scores?



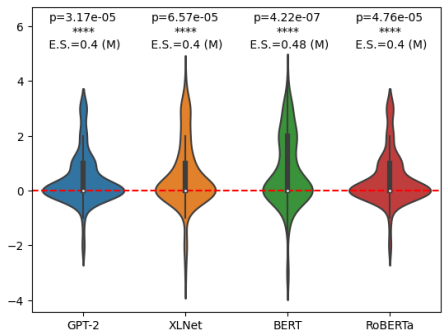
(a) GPT-2 (b) BERT (c) DeBERTa

Figure 9: Correlation between adjective frequencies and total sentence surprisal scores (GPT-2, BERT) or entailment scores (DeBERTa).

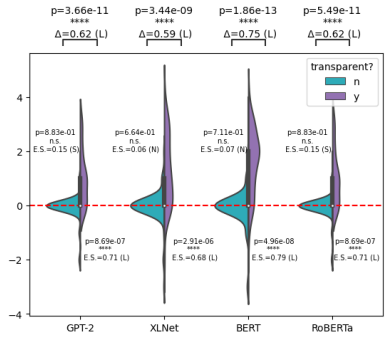
To answer these questions, we first computed, for each pair of adjectives, the differential number of tokens of A^- vs. A^+. Given that A^- is assumed to be overall more complex than A^+, the resulting differential number of tokens is expected to be positive. Figure 10 shows that this expectation is verified for all models, although the result seems to be driven by the transparent pairs only. This is not at all surprising given that tokenizers only have access to surface representations (strings) and as such cannot apply Büring's generalization to opaque pairs. We then tried to assess if differential numbers of tokens correlate with the surprisal contrasts measured on Task 1 for the two best-performing models (GPT-2, BERT); and the differential entailment scores measured on Task 3 for DeBERTa. The relevant scatter plots are shown in Figure 11 and suggest the existence of a weak positive correlation in the case of GPT-2, and a weak negative correlation in the case of DeBERTa. In other words, for GPT-2 the differential in complexity between A^+ and A^- tends to make the surprisal contrast between (11b) and (11a) bigger, which is somehow expected, while for DeBERTa, the differential in complexity between A^+ and A^- tends to make the contrast in entailment strength between (14a) and (14b) smaller, which is unexpected. Differential numbers of tokens however, are perhaps not extremely informative if the parses generated by the tokenizers do not match the morphology of their input in the first place.

For that reason, we assessed how accurate the tokenizers were in segmenting the adjectives from our dataset according to their actual morphological decomposition. In the general case, tokenizers managed to get the right parses between 42 and 48% of the time, but the accuracy significantly dropped when focusing on adjectives with plurimorphemic parses: GPT-2 and RoBERTa (which rely on the same tokenizer) only achieved a 4% accuracy, while BERT and XLNet respectively achieved 12 and 15%. Finally, we assessed how often tokenizations of morphologically transparent negative adjectives from our dataset involved a boundary between

the negative morpheme and the base, since this decomposition is in theory the source of the complexity difference between positive and negative adjectives. We found that GPT-2 and RoBERTa only had a 21% accuracy in this particular task, while XLNet and BERT achieved a 60% accuracy. Those overall poor results imply that, even though some models exhibit the expected differences in complexity between A^+ and A^- in transparent pairs, and somehow rely on those differences to derive surprisal contrasts (in the case of GPT-2), they start out with representations that do not match linguistic theory.

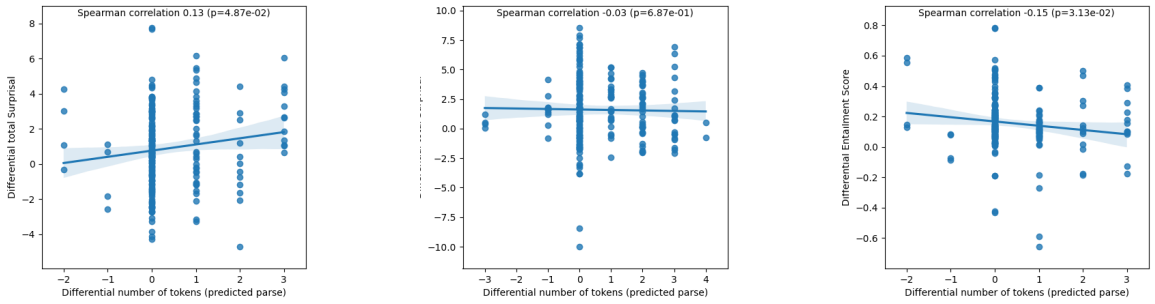


(a) All groups. Same tests and corrections as in previous tasks.



(b) T- vs. O-group. Same tests and corrections as in previous tasks.

Figure 10: Differential number of tokens between A^- and A^+.



(a) GPT-2 (b) BERT (c) DeBERTa

Figure 11: Correlation between differential number of tokens and differential surprisal scores (GPT-2, BERT) or entailment scores (DeBERTa).

5. Conclusion

We assessed various LLMs on their interpretation of antonymic adjectives and their respective negations, in particular, with regards to the Inference Towards the Antonym, which is expected to be stronger for negated *positive* adjectives as opposed to negated *negative* adjectives, and even more so for morphologically transparent pairs. Using measures of surprisal (Task 1), probabilities of entailment (Task 2) and vector similarities (Task 3), we found some evidence that two basic models (BERT and GPT-2), and one model fine-tuned for Natural Language Inference (DeBERTa) captured the predicted polarity contrast, and, in some cases, the magnifying effect of morphological transparency. More "advanced" models (on regular benchmarks)

noticeably performed lass well on the tasks at stake. More targeted analyses however, showed that some reasonable expectations about the models' behavior were not met. In Task 1, even the LLMs which managed to give human-like "judgments" on minimal pairs, did not seem to focus on the right individual words to produce them and/or seemed to overly rely on bare adjective frequencies. In Task 2, we reported mixed results and, even for the best-performing model, observed correlations between entailment scores and frequencies of the target adjectives, as well as between differential numbers of token and differential entailment scores–which implies that the model might have relied on superficial cues to draw its conclusions. In Task 3, even when the LLMs' contextual representations appeared to capture ITA-related topological inequalities, the very same spaces were characterized by the stronger, very much unexpected topological regularity consisting in a clustering of bare antonyms on the one hand, and their negations, on the other hand. This clustering moreover seemed to depend on the number of tokens within each adjective.

In sum, some LLMs seem to be shallowly accurate in their treatment of antonymic adjectives, but also deeply "confused" about the sources of the relevant contrasts. More generally perhaps, this study questions how LLMs (and, in retrospect, humans!) can be sensitive to concepts such as markedness, and pragmatic competition. Should markedness be identified with formal complexity, and should differences in word frequencies be seen as the consequence of differences in markedness? Should the definition hold in the opposite direction? Or should markedness be seen as the result of an *interaction* between complexity and typicality? Finally, regarding competition and the nature of alternatives, it is worth nothing that the pragmatic framework we used makes the assumption that antonymic adjectives interact within a fixed *pair* but in practice, the negation of a given positive adjective might compete with more than one negative counterpart, and vice versa. This might make an account of the ITA more challenging, in that the *number* and relevance of potential competitors to a given negated adjective, in addition to the differential of complexity contributed by each competitor, might eventually play a role in the mitigation effect observed.

References

Aina, L., R. Bernardi, and R. Fernández (2019, June). Negated adjectives and antonyms in distributional semantics: not similar? *Italian Journal of Computational Linguistics 5*(1), 57–71.

Bender, E. M. and A. Koller (2020, July). Climbing towards NLU: On meaning, form, and understanding in the age of data. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, pp. 5185–5198. Association for Computational Linguistics.

Bierwisch, M. (1989). *The Semantics of Gradation*, pp. 71–261. Springer Berlin Heidelberg.

Brants, Thorsten and Franz, Alex (2006). Web 1t 5-gram version 1.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and

D. Amodei (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Eds.), *Advances in Neural Information Processing Systems*, Volume 33, pp. 1877–1901. Curran Associates, Inc.

Büring, D. (2007, October). Cross-polar nomalies. *Semantics and Linguistic Theory 17*, 37.

Büring, D. (2007). More or less. In *Proceedings of the 43th Annual Meeting of the Chicago Linguistic Society*.

Charles, W. G. and G. A. Miller (1989). Contexts of antonymous adjectives. *Applied Psycholinguistics 10*(3), 357–375.

Cong, Y. (2022, July). Pre-trained language models' interpretation of evaluativity implicature: Evidence from gradable adjectives usage in context. In V. Pyatkin, D. Fried, and T. Anthonio (Eds.), *Proceedings of the Second Workshop on Understanding Implicit and Underspecified Language*, Seattle, USA, pp. 1–7. Association for Computational Linguistics.

Devlin, J., M. Chang, K. Lee, and K. Toutanova (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR abs/1810.04805*.

Futrell, R., E. Wilcox, T. Morita, P. Qian, M. Ballesteros, and R. Levy (2019, June). Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, pp. 32–42. Association for Computational Linguistics.

Gotzner, N., S. Solt, and A. Benz (2018, November). Adjectival scales and three types of implicature. *Semantics and Linguistic Theory 28*, 409.

Grand, G., I. A. Blank, F. Pereira, and E. Fedorenko (2022, April). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature Human Behaviour 6*(7), 975–987.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001 - NAACL '01*. Association for Computational Linguistics.

Harris, Z. S. (1954). Distributional structure. *Word 10*(2-3), 146–162.

He, P., X. Liu, J. Gao, and W. Chen (2020). Deberta: Decoding-enhanced BERT with disentangled attention. *CoRR abs/2006.03654*.

Horn, L. R. (1989). *A Natural History of Negation*. Chicago, IL: University of Chicago Press.

Jeretic, P., A. Warstadt, S. Bhooshan, and A. Williams (2020, July). Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, pp. 8690–8705. Association for Computational Linguistics.

Justeson, J. S. and S. M. Katz (1991). Co-occurrences of antonymous adjectives and their contexts. *Computational Linguistics 17*(1), 1–20.

Krifka, M. (2007). Negated antonyms: Creating and filling the gap. In U. Sauerland and P. Stateva (Eds.), *Presupposition and Implicature in Compositional Semantics*, pp. 163–177. London: Palgrave Macmillan UK.

Levy, R. (2008, March). Expectation-based syntactic comprehension. *Cognition 106*(3), 1126–1177.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint*

arXiv:1907.11692.

Misra, K. (2022). minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

Radford, A., K. Narasimhan, T. Salimans, and I. Sutskever (2018). Improving language understanding by generative pre-training.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever (2019). Language models are unsupervised multitask learners.

Rett, J. (2015). *The semantics of evaluativity*. Oxford: Oxford University Press.

Ruytenbeek, N., S. Verheyen, and B. Spector (2017, October). Asymmetric inference towards the antonym: Experiments into the polarity and morphology of negated adjectives. *Glossa: a journal of general linguistics 2*(1).

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin (2017). Attention is all you need. *CoRR abs/1706.03762*.

Wilcox, E., R. Levy, T. Morita, and R. Futrell (2018, November). What do RNN language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Brussels, Belgium, pp. 211–221. Association for Computational Linguistics.

Wilcox, E. G., R. Futrell, and R. Levy (2023, 04). Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*, 1–44.

Williams, A., N. Nangia, and S. Bowman (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122. Association for Computational Linguistics.

Yang, Z., Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov, and Q. V. Le (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 5754–5764.

Exclusivity and exhaustivity of disjunction(s): a cross-linguistic study.[1]

Andreea C. NICOLAE — *Leibniz-Centre General Linguistics*
Aliona PETRENCO — *Humboldt University*
Anastasia TSILIA — *Massachusetts Institute of Technology*
Paul MARTY — *University of Malta*

Abstract. Most natural languages have more than one linguistic form available to express disjunction. One of these forms is often reported by native speakers to be more exclusive than the other(s) and, in recent years, it has been claimed that some languages may in fact have dedicated exclusive disjunctions. In this paper, we report on a series of verification studies investigating the robustness of the exclusivity inference associated with different disjunction markers within and across five different languages and extend this investigation to another, related type of inference, namely the exhaustivity inference. In our results, we found that complex disjunctions were generally more likely to be interpreted exclusively than simplex ones and that, in some languages, further differences exist among the complex disjunctions. Exhaustivity inferences were much less robust and, by contrast, showed little-to-no difference among disjunction types. We lay out possible directions for interpreting these results.

Keywords: disjunction, exclusivity, exhaustivity, complexity, alternatives, cross-linguistic semantics

1. Introduction

Disjunctive sentences like (1) are ambiguous between an inclusive and an exclusive interpretation. Most, if not all extant accounts of this phenomenon assume that plain disjunctions like English *or* encode an inclusive meaning, yielding the literal interpretation in (1a). In positive sentences like (1), this inclusive meaning can be strengthened to an exclusive one via scalar inferencing, yielding the enriched interpretation in (1b).

(1) Asher will order beer or wine.
   a. Asher will order beer or wine (possibly both). *Inclusive*
   b. Asher will order beer or wine, but not both. *Exclusive*

English, like many other languages, has yet another way of expressing disjunction: in addition to plain *or*, we also find the morphologically complex disjunction *either*. . . *or*; similarly, in German, we find a plain disjunction, *oder*, and a more complex one, *entweder*. . . *oder*. Many languages show in fact a three-way and even four-way distinction, with multiple simplex and complex disjunctive forms available. For instance, in Russian, we find *ili*, *ili*. . . *ili* and *libo*. . . *libo*, in Hungarian *vagy*, *vagy*. . . *vagy* and *akár*. . . *akár*, in French *ou*, *ou*. . . *ou*, *ou bien*. . . *ou bien* and *soit*. . . *soit*, in Romanian *sau*, *ori*, *ori*. . . *ori* and *fie*. . . *fie*. The multiplicity of disjunctive particles in these languages raises an immediate question: do all these particles convey the same meaning and if not, what are the dimensions of variation?

[1]The authors are grateful to the audience at Sinn und Bedeutung in Bochum, the semantics colloquium in Göttingen, the workshop on Logic, Grammar and Meaning in Milan, the Brown Bag lunch series in Berlin and the Nihil workshop in Amsterdam, where different incarnations of this material has been presented. We are particularly thankful to Maria Aloni, Nina Haslinger, Clemens Mayr, Uli Sauerland, Viola Schmitt and Yasu Sudo for their invaluable feedback. This research was supported by the DFG grant NI-1850/2-1, awarded to A. Nicolae.

 In: Baumann, Geraldine, Daniel Gutzmann, Jonas Koopman, Kristina Liefke , Agata Renans, and Tatjana Scheffler (eds.) 2024. Proceedings of Sinn und Bedeutung 28. Bochum: Ruhr-University Bochum, 1098-1113.

An intuition commonly reported in both the expert and non-expert literature is that the different forms available for expressing disjunction within a language relate to the extent to which they associate with an exclusive interpretation. Typically, authors of logic textbooks use the more complex disjunction(s) to exemplify the meaning of the logical exclusive operator XOR, in line with the paraphrases that linguistically naive speakers often provide for these complex forms. Similar intuitions are found in the expert literature where, for some languages, complex disjunctions have been claimed to unambiguously convey an exclusive interpretation. For instance, Spector (2014: p.13-18) claims that, in French, the reiterated disjunction *soit... soit*, unlike the simplex disjunction *ou*, obligatorily gives rise to an exclusive inference like the one in (1b) in non-embedded contexts. Szabolcsi (2015: p.194-197) extends this claim to other disjunctions with reiterated particles such as French *ou... ou*, Russian *ili... ili* or Hungarian *vagy... vagy*.[2] Thus, according to the literature, some languages ought to have dedicated 'exclusive' disjunctions which obligatorily trigger the (otherwise optional) exclusive inference associated with disjunction. Whether or not this claim is empirically correct remains an open question which has not been systematically investigated across languages using quantitative methods.

A weaker claim, closer to what Spector (2014) ultimately endorses, is that complex disjunctions obligatorily trigger strengthening, but that this strengthening need not be to exclusivity. In support of this claim, Spector observes for instance that sentences like (2), where the complex disjunction *soit... soit* is embedded in the scope of a universal quantifier, need not yield the strong exclusivity inference in (2a); rather, it can yield the weaker inference in (2b), which leaves open the possibility that some guests ordered both beer and wine.

(2) Chaque invité a pris soit de la bière soit du vin.
'Every guest ordered SOIT beer SOIT wine'
a. Every guest ordered one or the other but not both.
b. It's not the case that every guest ordered both.

Building on this observation, we can then ask whether this weaker claim generally holds of complex disjunctions. That is, do complex disjunctions generally give rise to strengthened meanings, irrespective of the nature of the strengthened meaning? Note that, while Spector's observation pertains to the occurrence of *soit... soit* in the scope of a universal quantifier, the claim of interest extends in theory to unembedded environments, raising the question of whether the use of complex disjunctions in these environments may force other forms of non-exclusive enrichment. To answer this question, we thus need to consider other inferences generally associated with unembedded disjunctions. One such inference is the *exhaustivity* inference.[3] This inference, less extensively discussed in the literature on disjunction, makes reference to relevant alternatives to the mentioned disjuncts, rather than to their conjunction, and

[2]Szabolcsi (2015: p.197) also claims that, in this regard, reiterated complex disjunctions should be distinguished from non-reiterated complex disjunctions like English *either... or*, which, she argues, retain both their inclusive and exclusive flavors. As far as we know, this claim has not yet been put to the test.

[3]Another prominent type of inferences associated with unembedded disjunction are ignorance inference, e.g., the inference from (1) that the speaker doesn't know which of the two Asher ordered (i.e., both disjuncts are living possibilities in the speaker's mind). For space reason, we do not discuss these inferences in this paper; in fact, the verification studies we report on below were specifically designed to factor out the potential effect of these inferences on participants' judgments. We refer the reader to Degano et al. 2023 for a recent experimental investigation of ignorance inferences and to Nicolae 2017 for an argument that these inferences should count towards a requirement of obligatory strengthening.

says that they are not true (Gotzner et al., 2020). In the case of the sentence in (1), repeated below in (3), this inference says Asher will not order anything *besides* beer or wine, as illustrated in (3a). Importantly, we note that this inference does not carry any commitment as to whether he will order both drinks, namely the exclusivity component. Similarly, the exclusive inference does not carry any commitment as to whether Asher won't order anything besides beer or wine.

(3) Asher will order beer or wine.
   a. Asher will order beer or wine and nothing else. *Exhaustive*
   b. Asher will order beer or wine, but not both. *Exclusive*

In the remainder of the paper, we present a series of verification studies investigating the robustness of the exclusivity and exhaustivity inferences associated with different disjunction markers within and across five different languages.

2. Experiments

In the following, we present a series of studies investigating the robustness of different inferences across different disjunction markers, both within and across five languages: English, French, Romanian, Russian and Greek. For each language, we chose three of the most commonly used disjunctive markers, one simplex and two complex (with the sole of exception being English which only employs one type of complex disjunction). The two inferences under investigation were the exclusivity and the exhaustivity (ad-hoc) inferences. The three factors – language, disjunction type and inference type – were manipulated between-subjects. The disjunctive constructions tested in each language are schematically described in Table 1.

	D1	D2	D3
English	A *or* B	*either* A *or* B	n/a
French	A *ou* B	*ou bien* A *ou bien* B	*soit* A *soit* B
Romanian	A *sau* B	*fie* A *fie* B	*ori* A *ori* B
Russian	A *ili* B	*libo* A *libo* B	*ili* A *ili* B
Greek	A *i* B	*i* A *i* B	*ite* A *ite* B

Table 1: Disjunctive constructions tested in all five languages; D1 are simplex disjunctions whereas D2 and D3 are all complex.

2.1. Participants

Participants were recruited online using Prolific (minimum prior approval rate: 90%; nationality, country of birth and first language were controlled for, depending on the language being tested). Participants were paid approximately £1.7 for their participation (£8/hr). In total, 564 subjects took part in the Exclusivity studies and 533 in the Exhaustivity studies (see details in Table 2), yielding between 30 and 45 subjects per disjunction in each group. All participants gave written informed consent prior to experimentation. All data were collected and stored in accordance with the provisions of Data Protection Act 2018.

	Exclusivity	Exhaustivity
English	90	89
French	127	119
Romanian	111	106
Russian	107	94
Greek	129	128
Total	564	533

Table 2: Number of participants recruited for both sets of studies.

2.2. Materials and design

The experiments were run as online surveys. At the beginning of the survey, participants were given general instructions (translated by native speakers into the corresponding languages). They were told that they would witness a guessing game between two friends, Kate and Henry. The game was described as follows:

> *Instructions* – Kate and Henry are two friends who like playing games. In this experiment you will witness one of their games. The rules are as follows: Kate draws two pictures and doesn't show them to Henry. The first picture depicts a situation and comes with a sentence describing it. The second picture depicts a follow-up scene. She shows Henry the first picture, depicting the situation, and asks him to make a guess about what's going to happen. Then, Kate presents the second picture with the follow-up scene. Your task will be to judge whether Henry's guess was right by clicking the 'yes' or 'no' button.

Each trial consisted of a scenario unfolding over three scenes, where the test sentences appeared in the second scene. The structure of the scenarios was the same across all trials: the first scene set the stage of a story by displaying a picture together with a short sentence describing a future event; the second scene showed a character making a guess about what was going to happen next in that story in relation to the relevant event; finally, the third and last scene revealed the outcome of the story by means of a novel picture accompanied by the lead-up 'Here's what happened'. The participants moved from one scene onto the next by clicking a button at the bottom of the page; the picture(s) from the previous scene(s) remained on the page as the scenario progressed, such that the final scene consisted of all 3 pictures, as shown in Figure 1.

In the test trials, the character's guesses involved disjunctive sentences of the form *[Pronoun] will [verb] [(D) A D B]* such as *She will bring (either) a bouquet or balloons*. The [pronoun] term always agreed with the subject of the sentence displayed in the first scene; the [verb] term was an action verb; the disjunctive phrase [(D) A D B] involved a simplex or complex disjunction type connecting two common nouns (A, B) denoting inanimate, concrete objects.

Test sentences were presented with one of three possible story outcomes obtained by manipulating the contents of the final scene picture; these constituted the three conditions of interest, namely TRUE, FALSE and TARGET. The TRUE and FALSE conditions were constant across experiments, while the TARGET condition differed. In the TRUE and FALSE conditions, the final scene made the disjunctive sentences true and false, respectively, independently of the type of

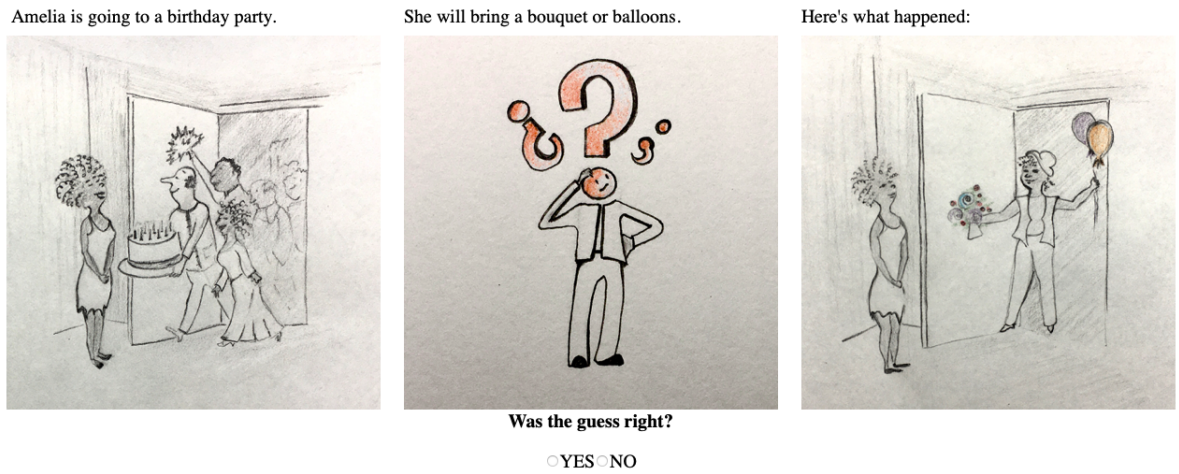


Figure 1: Example of scenario used in the TARGET trial for the Exclusivity studies. Scenarios unfolded before the participants, one scene at a time.

inference being tested; this was achieved by making the disjunction true via the truth of one of the disjuncts, or false via the falsity of both disjunct (see details in (4)). The TARGET condition, on the other hand, differed across the two experiments since it varied according to the inference type being tested.[4] In the Exclusivity studies, which tested for presence of the exclusivity inference, both objects mentioned in the guess appeared in the final image. Such an outcome made the test disjunctive sentence false if the exclusivity inference was present (expected answer: 'No'), but true if it was absent (expected answer: 'Yes'). In the Exhaustivity studies , which tested for the presence of the ad-hoc/exhaustivity inference, only one of the objects mentioned in the guess appeared in the final image, but crucially also an additional, unmentioned, object. Such an outcome made the test disjunctive sentence false if the exhaustive inference was present (i.e., participants would be expected to select 'No' to the question), but true if it was absent (i.e., participants would be expected to select 'Yes' to the question). Note that by only presenting one of the two objects mentioned in the disjunctive sentence we avoided having participants judge the sentence based on its exclusivity inference potential.

(4) Possible outcomes for target sentences *'(either) A or B'*
   a. TRUE: A
   b. FALSE: C
   c. TARGET$_{exclusive}$: A and B
   d. TARGET$_{exhaustive}$: A and C

Target sentences were tested in all three conditions, with three iterations of each condition, yielding 9 test items. 18 non-disjunctive filler items were added in order to make the target items less visible across the experiment: 6 true, 6 false and 6 open to interpretation. Participants started the experiment with two practice trials and then completed the 27 test trials, presented to them in a randomized order.

[4]Two of the three target items had to be changed in the Exhaustivity studies due to the lack of easily accessible and salient third alternatives beyond the two mentioned in the target sentence. The Exhaustivity studies crucially relied on there being such alternatives to the disjuncts, something that the target items in the Exclusivity studies didn't necessitate.

Inference type was a between-subject factor such that no participant saw a test sentence in both types of target conditions shown in (4). This constitutes the only difference between the Exclusivity and the Exhaustivity studies. Within each experiment, disjunction type was also manipulated and this too was a between-subject factor. This was done so as not to encourage implicit, comparative judgments between disjunctive constructions.[5] All materials created for the English version were adapted and translated into French, Romanian, Russian and Greek by linguistically-trained native speakers.

2.3. Data preparation

Data preparation and analysis were carried out in the R statistical environment (R Core Team, 2023) using the Hmisc (Harrell, 2023), lme4 (Bates et al., 2015), and car (Fox and Weisberg, 2019) packages for the R statistics program. Responses from 37 subjects in the Exclusivity studies (6.5% of the sample) and from 43 subjects in the Exhaustivity studies (8% of the sample) were removed prior to analyses because their performance to TRUE and FALSE controls did not reach the pre-established threshold of 80% accuracy.

2.4. Results

Responses to the test items are summarized in Figure 2. In both experiments, the rate of 'No' responses (i.e., 'wrong guess') was lowest in the TRUE conditions (all Ms$<$ 5%), highest in the FALSE conditions (all Ms$>$ 90%) and somewhat intermediate in the TARGET conditions. Recall that, in the TARGET conditions, this measure stands proxy for the rate of exclusive/exhaustive interpretations, meaning that the higher the rate of 'wrong guess' responses, the more exclusive/exhaustive inferences being drawn. In our statistical analyses, we assessed, for each inference type in each language, whether responses in the TARGET conditions differ as a function of the disjunction type; we report the results of these analyses below.

2.4.1. Exclusivity studies

In the TARGET conditions, every disjunction in the five languages tested received an intermediate rejection rate, i.e., in-between those observed for their TRUE and FALSE baselines. These results are expected only if the disjunctions of interest are assumed to be ambiguous between an inclusive and an exclusive reading. The mean rejection rates for D2 and D3 were relatively uniform across languages, with 8 out of 9 instances in the 60-75% range, while the rates for D1 showed more variations, ranging from 20% in Romanian to 54% in Greek.

For each language, we fitted a GLMER model (logit link function), predicting responses in the TARGET conditions from the fixed effect of disjunction (dummy coded). Each model included by-participant and by-item random variance for the intercept, which was the maximal random effect structure supported by the data.[6] Each of these models was compared to a null model missing only the fixed effect of disjunction. The model with the fixed effect of disjunction was

[5]Nicolae and Sauerland (2016) have shown that speakers' judgements of exclusivity are affected when presented with multiple disjunction markers within the same experiment.

[6]The model for French triggered a singular fit warning due to the by-item random variance for the intercept being estimated very near zero. As a sanity check, this model was refitted without the random intercept for items. The values of the coefficients of the refitted model were the same as before. While this warning only arose for this model, we note that the estimated variance for the item random effect was relatively small in all models.

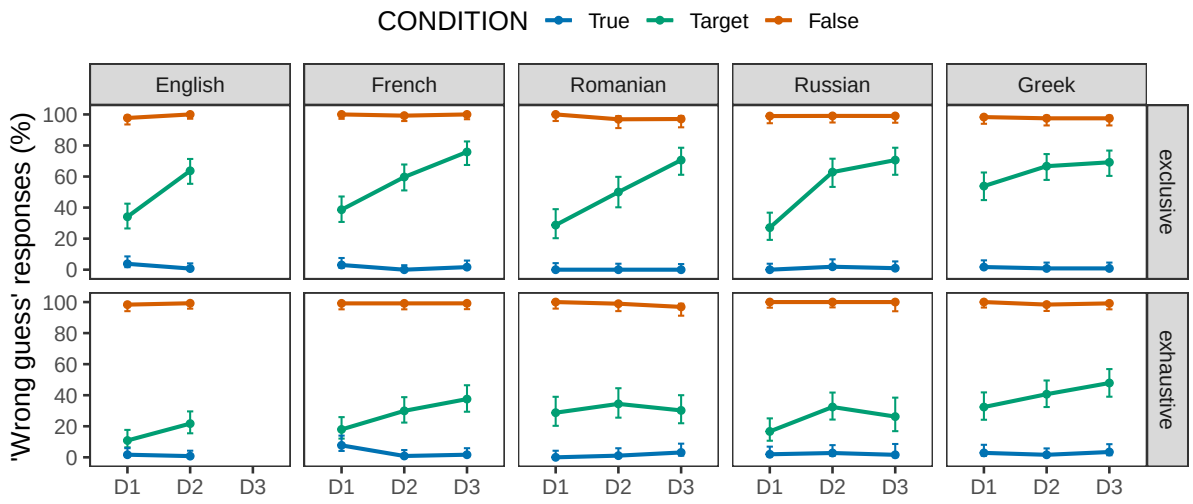


Figure 2: Mean rejection rate (i.e., proportion of 'wrong guess' responses) to the test trials by inference type, language, disjunctive marker and picture condition. Error bars represent 95% binomial confidence intervals.

found to provide a significantly better fit to the data compared to the null model for English ($\chi^2_1 = 25.26$, $p < .001$), French ($\chi^2_2 = 21.06$, $p < .001$), Romanian ($\chi^2_2 = 34.64$, $p < .001$), Russian ($\chi^2_2 = 36.51$, $p < .001$) but not for Greek ($\chi^2_2 = 4.64$, $p = 0.09$), where the mean rejection rate for D1 was only marginally lower than those for D2 and D3. In all other languages with reiterated disjunctions (French, Romanian and Russian), both D2 and D3 yielded significantly higher rejection rates than D1 (all βs> 3.27, all ps$< .05$). Further reliable contrasts were found between D2 and D3 in French ($\beta = 2.55$, $p = .05$) and Romanian ($\beta = 7.23$, $p < .05$), showing that distinct reiterated disjunctions in these languages prompt exclusive interpretations to a different extent. No such contrast was found in Russian ($\beta = 0.59$, $p = 0.7$).

2.4.2. Exhaustivity studies

All disjunctions received an intermediate rejection rate in the TARGET conditions, except for the simplex disjunctions in English and French. Nevertheless, the mean rejection rates in these conditions were relatively low across languages, with 13 out 14 instances in the 10-40% range (Greek D3: 48%). Thus, as can be seen in Figure 2, the rate to which exhaustive inferences were drawn was lower than the rate to which exclusive inferences were drawn across the board.

As before, for each language we fitted a GLMER model predicting responses in the TARGET conditions from the fixed effect of disjunction. Each model included by-participant and by-item random variance for the intercept.[7] Each of these models was compared to a null model missing only the fixed effect of disjunction. The model with the fixed effect of disjunction was found to provide a significantly better fit to the data compared to the null model only for English and English only (English: $\chi^2_1 = 5.12$, $p < .05$; French: $\chi^2_2 = 5.38$, $p = .07$; Romanian: $\chi^2_2 = 0.97$, $p = .61$; Russian: $\chi^2_2 = 2.66$, $p = .26$; Greek: $\chi^2_2 = 3.39$, $p = 0.18$).

[7]As in the Exclusivity studies, we ran into a singular fit warning due to the by-item random variance for the intercept being estimated at zero. In this experiment, this was the case for both French and Romanian.

3. Discussion

The findings of these experiments can be summarized as follows:

- All the disjunctions tested in this study yielded ambiguity patterns showing that, no matter how 'exclusive' they feel, they all allow an inclusive interpretation.
- Complex disjunctions generally yielded higher rates of exclusive interpretations than simplex ones across languages.
- Speakers' propensity to interpret a disjunction exclusively varies substantially: (i) there is wide cross-linguistic variation in how exclusive simplex disjunctions are interpreted (e.g., Romanian vs. Greek), and (ii) further contrasts may exist among complex disjunctions within the same language (e.g., in French and Romanian).
- Exhaustivity (ad-hoc) inferences arose cross-linguistically but were much less derived than exclusivity inferences.

In the remainder of this section we will discuss the implications of our results for current theories of implicatures.

3.1. Comparison of inference strengths: exhaustive versus exclusive

We begin with a short discussion of the comparison between exclusive and exhaustive inferences. To reiterate, our results indicate that exhaustive inferences were derived less often than exclusive inferences in our studies. We believe that this finding can be explained in reference to the nature of the different types of alternative involved in the derivation of these inferences. Consider again the example from earlier, repeated below in (5). Deriving the exhaustive inference associated with this sentence requires alternatives like those in (5a), all of which involve generating ad-hoc competitors to the disjuncts, i.e., competitors constructed from contextual, rather than conventional linguistic factors. As it is easy to verify, these alternatives can be negated altogether without giving rise to a contradiction, yielding the inference in (5b).

(5) Asher will order (either) beer or wine.

a. *Alternatives:* { Asher will order [lemonade]. / Asher will order [whiskey]. / Asher will order beer and [lemonade]. / ⋮ }

b. *Exhaustive inference:* Asher will order nothing else besides beer and wine.

Exclusivity inferences, on the other hand, are generally assumed to arise from the more basic lexical competition between 'or' and its scalemate, the logically stronger connective 'and'. In this case then, the competitors of interest need not be set up by the context for the competition to arise: this competition directly arises due to the conventional semantic content of the relevant connectives. Thus, the results we obtained could be a by-product of this difference in the make-up of both types of inference. Specifically, it is possible that the set-up of our studies made it so that constructing novel ad hoc competitors to the disjuncts on trial-to-trial basis was far more demanding than simply retrieving the invariant lexical competitor to the disjunctive marker, hence the lower rates of exhaustivity inference that we observed.

3.2. Variation in inference strength of exclusivity

Our findings disconfirm the claim that reiterated disjunctions in languages like French, Russian or Romanian are dedicated 'exclusive' disjunctions categorically distinct from simplex ones. Crucially, however, these findings remain largely in line with the layman's intuition and support the weaker claim that complex disjunctions are more strongly associated with an exclusive interpretation than simplex ones. In the following we will offer some thoughts on what could be driving this tendency and how we might begin to formalize such contrasts.

In principle, this cross-linguistic tendency to interpret complex disjunctions exclusively more so than simplex disjunctions could be explained in reference to cost-benefit principles like Horn's 1984 *R Principle*: since it would be more economical for speakers to use a simpler form to convey the literal, inclusive meaning of disjunction, the use of a more complex disjunction can be taken as signaling the intent of the speaker to depart somehow from that literal meaning, e.g., to convey the enriched, exclusive meaning. This is, in fact, what laid the groundwork for the proposal put forward in Nicolae and Sauerland 2016 (henceforth N&S).

On N&S's proposal, simplex and complex disjunctions compete with each other. Their proposal is motivated by the finding that, when presented with both *or* and *either or* (or *oder* and *entweder oder* in the German variant) in the same experimental session, participants rated the complex disjunction as more exclusive than the simplex one, whereas no such contrast was observed when the two forms were tested in isolation.[8] The crux of their proposal is that the simplex disjunction does not compete with conjunction but rather with the complex disjunction, which itself receives its strengthened meaning via competition with conjunction. Crucial to their account is the assumption that assertively used sentences contain not only an exhaustification operator, but also a covert doxastic operator which is adjoined at LF (cf. Meyer (2013); see also also Kratzer and Shimoyama (2002), Chierchia (2006) and Alonso-Ovalle and Menéndez-Benito (2010) for related proposals). This operator, generally referred to as the K-operator, can be thought of as the necessity modal, with the semantics in (6) (following Gazdar (1979), a.o.).

(6) $[\![\Box_x \text{ p}]\!] = \lambda w.\forall w' \in \text{Dox}(x)(w) : p(w')$
$w' \in Dox(x)(w)$ iff given the beliefs of x in w, w' could be the actual world.

Given this operator, as well as the exhaustification operator *exh* responsible for deriving scalar implicatures (Chierchia et al., 2012), N&S propose the following LF for *either. . . or*:

(7) **LF for *either–or*:** $\Box exh[p \vee q]$ (N&S: ex. 21)
a. $Alt(p \vee q) = \{p \vee q, p \wedge q\}$
b. $[\![\Box exh[p \vee q]]\!] = \Box[p \vee q] \wedge \Box\neg[p \wedge q]$

Assuming the meaning above for *either–or*, they propose that *or* takes as its alternative this stronger meaning under the LF in (8), delivering the weaker meaning in (8b).

(8) **LF for *or*:** $exh\Box[p \vee q]$ (N&S: ex. 22)
a. $Alt(\Box[p \vee q]) = \{\Box[p \vee q], \Box exh[p \vee q]\}$
$= \{\Box[p \vee q], \Box[p \vee q] \wedge \Box\neg[p \wedge q]\}$

[8] There were two experiments per language, and each involved giving ratings on a 7-point scale; in one experiment, participants had to judge the extent to which a disjunctive sentence *A or B* suggests *not A and B*; in the other, they had to judge whether one could conclude *only one* given the disjunctive statement.

b. $[\![exh\Box[p \vee q]]\!] = \Box[p \vee q] \wedge \neg[\Box(p \vee q) \wedge \Box\neg(p \wedge q)]$
$= \Box[p \vee q] \wedge \neg\Box\neg[p \wedge q]$

This proposal can account for the simplex-complex two-way distinction, especially in experimental setups where the two forms are pinned against one another, like the ones which N&S aim to account for. When it comes to setups like the one in the present study, where the *or/either or* contrast was between- rather than within-participants, the idea that *or* would be strengthened with respect to *either... or* rather than *and* becomes less appealing. One point against it comes from the observation that alternatives which are structurally more complex than the asserted sentence are generally not considered when calculating implicatures, unless the particular linguistic structure has been made salient in the discourse (Katzir, 2007). Since participants in the simplex conditions were not presented with the complex variant(s) during the survey, the argument goes that they should not have been able to strengthen *or* (and its cross-linguistic equivalents) via negation of the complex disjunction *either... or* (and its equivalents); in other words, participants in the simplex disjunction condition would only be expected to strengthen via negation of the stronger conjunctive alternative. The question thus remains, why are participants more likely to interpret a complex disjunction as exclusive than a simplex one?[9]

In our attempt to better understand what might be behind this difference in robustness between different disjunction markers, consider the finding from van Tiel et al. (2016) (building on Baker et al. 2009; Doran et al. 2012) that some scalar elements are more likely to give rise to a scalar implicature than others, with *cheap/free, sometimes/always, some/all, possible/certain* being at the high-end of the strength scale and *ugly/hideous, silly/ridiculous, tired/exhausted, content/happy* at the low-end with fewest scalar inferences being drawn. Among other factors, van Tiel et al. (2016) show that (part of) the variability observed is predicated by the boundedness of the scalemate involved, namely whether or not it corresponds to an end-of-scale expression, i.e., given a lexical scale $<\alpha, \beta>$, the distinctness of α and β is greater if β denotes an end point on the dimension over which it quantifies.

So could boundedness also explain the contrasts that we observed? One obvious concern here is that, in our case, we are dealing with the same inference, at least superficially, derived from the use of two logically equivalent elements. However, if the story advocated for by N&S is to be adopted, and the two scalar items, *or* and *either... or*, appeal to different alternatives, then boundedness might be a relevant notion afterall. Since *either... or* has *and* as an alternative, an end-of-scale expression, whereas *or* has *either... or* as an alternative, which is not bounded as far as its linguistic meaning goes, the fact that *either or* triggers a stronger scalar implicature than *or* can be explained by the account put forward by van Tiel et al. (2016). Nevertheless, as discussed above, we believe that the account in N&S does not readily extend to the experimental design we employed. Thus, in the remainder of this paper, we would like to sketch two alternative accounts, which may ultimately prove to be related to one another.

3.2.1. Cues to local exhaustification

The first account revolves around cues to local exhaustification. It builds on the observation that complex disjunctions usually facilitate, if not favor, a contrastive focus configuration. Thus

[9]We note here that N&S's account also falls short of an explanation as to why complex disjunctions may be perceived as more exhaustive than simplex disjunctions since their account crucially builds on the interaction between the *and*-alternative and the two disjunctive forms.

for instance, in declarative sentences, English *or* does not easily allow focus on each individual disjunct, unlike *either or*, as exemplified by the contrast in (9).[10]

(9) a. ??ASHER$_F$ or BILL$_F$ will visit Paris.
b. Either ASHER$_F$ or BILL$_F$ will visit Paris.

We propose that this configuration more readily calls for an interpretation where each disjunct is interpreted exhaustively, a reading along the lines of *Only Asher, or only Bill will visit Paris*. This can be achieved by taking *exh* to adjoin locally to each of the disjuncts, as in (10a)/(11a).[11] Depending on what the relevant alternatives are, notated here as subscripts on the respective *exh* operators, the result of this exhaustification process may yield the exclusive interpretation in (10b) or the exhaustive one in (11b) (or both).

(10) Complex: **Disj** A **Disj** B
a. $[exh_{\{A,B\}}(A) \vee exh_{\{A,B\}}(B)]$
b. $(A \wedge \neg B) \vee (B \wedge \neg A)$

(11) Complex: **Disj** A **Disj** B
a. $[exh_{\{A,C\}}(A) \vee exh_{\{B,C\}}(B)]$
b. $(A \wedge \neg C) \vee (B \wedge \neg C)$

On the assumption that contrastive focus in disjunction is a reliable cue to exhaustification, we would expect that disjunction involving narrow focus on the disjuncts should be associated with strengthened meanings more often than disjunction involving, say, broad focus. There are two possible ways of implementing this: (i) take *exh* to be optional and have its insertion be dependent on prosodic prominence, or (ii) take *exh* to be obligatory, and assume that prosodic prominence is associated with an increase in access to relevant alternatives. Such a proposal could even be taken a step further in order to account for differences among complex disjunctions. Specifically, we could argue that prosodic prominence is gradient and this gradience is an indicator of the inference strength. While this proposal is somewhat speculative, we believe that a production study looking into the prosodic prominence associated with different disjunction markers could be conducted to test this hypothesis.[12]

3.2.2. Cues to (levels of) uncertainty

Whereas the previous account was couched in terms of (strength of) cues to local exhaustification, the account we present in this section takes inference strength to correlate with variation in listener's certainty about the intended inference. Here too we identify two possible ways of couching this variation, and we discuss each of them in turn below.

[10] We note that this contrast is much less pronounced in post-verbal position, as in (i):
(i) a. Anushka will visit PARIS$_F$ or BERLIN$_F$.
b. Anushka will visit either PARIS$_F$ or BERLIN$_F$.
With simplex disjunctions, it is also possible to place the focus on the disjunctive marker itself, in which case the exclusive inference becomes quite strong. However, uttering a disjunctive statement with pitch accent on the disjunction only seems fully felicitous as a correction to a conjunctive statement.

[11] While we don't go into the details here, we do believe that the most likely underlying representation is one involving ellipsis, and thus clausal disjunction. Under this view then, the *exh* operator acts at the clausal level.

[12] There is currently a debate in the literature as to what should count when evaluating prosodic prominence which is why we remain agnostic.

On the Gricean approach to implicature calculation, a listener first considers relevant alternatives which the speaker could have uttered. In response to a weak utterance, a listener assumes that the speaker is uncertain about the truth value of stronger, more informative, alternatives, given than the speaker did not utter these. This step, on its own, only derives the weak inference (cf. *a primary implicature*) that the speaker is uncertain about the truth of stronger alternatives. It has been claimed, however, that a further step can be taken in order to derive the stronger inference of certainty regarding the falsity of stronger alternatives (cf. *a secondary implicature*). This step involves the additional assumption that the speaker is knowledgeable, or opinionated, with respect to the truth of alternative propositions (cf. *the epistemic step*) (Grice, 1967; Horn, 1972; Gazdar, 1979; Sauerland, 2004). It is not unnatural to suppose that the use of marked forms is meant to indicate a higher level of opinionatedness on the part of the speaker. This would then amount to higher rates of secondary implicatures for complex disjunctions. Note that within such an account, opinionatedness would be taken to be probabilistic. While this remains speculative for now, future studies could look into such possible correlations between perception of speaker expertise and rate of strengthening.[13]

The proposal we just sketched is neo-Gricean. Grammatical versions of this account handle the opinionantedness component via the K operator introduced previously in the context of N&S's proposal. Specifically, the distinction between primary and secondary implicatures is viewed as a scope interaction between K/□and *exh*, such that *exh*>□ delivers a primary implicature and □>*exh* a secondary one. We illustrated this point for the exclusivity inference(s) in (12).[14]

(12) a. **Strong exclusive interpretation:** $\Box\neg(p \wedge q)$
exclusive in every possible world under consideration.
b. **Weaker exclusive interpretation:** $\neg\Box(p \wedge q)$
exclusive in some of the possible worlds under consideration.

The two inferences above differ in terms of how strong the requirement for exclusivity is, with variation in strength being analyzed as a function of how many possible worlds satisfy the exclusivity requirement. Assuming that robustness of inference can be seen a reflection of strength of inference, as shown above, the problem is that this only gives us two levels of variation and it is unclear how it would be able to account for the three-way variation observed in languages like French and Romanian. A possible extension would be to appeal to a degree-based probabilistic semantics of modality, building on Swanson 2016; Yalcin 2007, 2010; Lassiter 2014, 2020; Moss 2015; Santorio and Romoli 2017.

The idea, in a nutshell, is to think of modals as measures of probability, thus allowing us to map propositions to a value on a probability scale. This would allow us to model the strength of the exclusivity inference in terms of the degree to which it is likely that the *not both* inference holds. Simplifying greatly, we can imagine that the covert modal posited for assertively-used sentences have variable strength, such that a disjunctive sentence can in fact be interpreted as conveying a degree, possibly not 100%, of certainty that the disjunctive statement holds.

[13] On this hypothesis, we could also imagine that disjunctions associated with a higher, more formal register (e.g. French *soit soit*) are also those more likely to give rise to exclusivity: if someone uses a disjunctive marker from a higher register, it gives the impression that they are more expert on the topic, hence more opinionated. We would like to thank Federica Longo for her suggestion to consider register as a relevant factor in modeling strength of implicatures.

[14] We assume the same system could be at play for ad-hoc inferences as well.

Assuming exhaustification can then proceed normally, the strengthened interpretation would amount to the interpretation in (13).[15]

(13) *n%* likely that *p or q* and *n%* likely not *p and q*

Depending on the strength of the modal, we could then envision it taking on different values depending on how strongly a given participant views the likelihood conveyed by the disjunctive statement, and in turn, the negation of its conjunctive inference. On this view, the higher the likelihood, the stronger the exclusivity inference. Do note that an issue with the account as presented so far is that it assigns the same *n%* to both the assertive and the implicated components, whereas intuitively it seems that the certainty level should only vary with respect to the implicated component, something that the neo-Gricean proposal presented above could capture. This issue is not insignificant but we nevertheless leave it as an open issue here.[16]

We have suggested that a possible implementation of the variability in strength of exclusivity could be achieved by adopting a degree-based probabilistic semantics of the covert modal operator.[17] We believe this can easily be extended to cases such as the ones in our experimental set-up which involved the future marker *will*, by re-analyzing it in terms of the speaker's belief in how likely a certain outcome is.

Summing up, the general line of reasoning we pursued here takes the disjunctive marker to affect what we take the speaker's epistemic state to be (albeit in ways we still don't fully understand) — be it because it modulates the strength of the K operator or because it modulates the likelihood of the opinionatedness assumption.

4. Concluding remarks

The results of our experiments showed that all disjunctions in the five languages we tested are ambiguous between an inclusive and an exclusive interpretation and that they may, but need not, differ in terms of how exclusive they are. These findings constitute a rebuttal of the categorical view whereby particular disjunctions are exclusive across the board. We sketched instead two non-categorical approaches that could explain the observed optionality in strength, building on the intuition that multiple aspects enter into the calculation, with prosodic marking and opinionatedness being two of the main factors we consider relevant. While in the previous section we discussed their roles in isolation, it is clear that they can work in concord, with prosodic marking being taken to relate to the activation of alternatives, and opinionantedness to the extent to which one can confidently exclude the activated alternatives.

Another question that arose from the experimental data presented here regards the multiplicity of disjunction particles. In languages like Russian and Greek, in contrast to languages like French and Romanian, we do not see gradual effects of exclusivity all the way, raising the following question: why would a language have three or more ways of expressing disjunction if only one or two gradients of exclusivity tend to be expressed? Could it be that differ-

[15]This interpretation assumes that *exh* occurs under the modal operator. It is not clear to us at this point how to tell apart the interpretations that arise from the two scope possibilities given this degree-semantics for the modal.

[16]But see Mandelkern and Dorst (2022) for a view that assertions should more generally be seen as weak.

[17]A similar approach could, in principle, also derive variability in strength of the exhaustivity inference. Given our results, however, any complete theory would need to also take into account the nature of the alternatives involved and the ease of retrieval.

ent disjunctions are responsible for different inferences? To begin answering this question, we tested another type of inference, namely the exhaustivity inference. This inference was overall much less robust than the exclusivity inference, and crucially, not remarkably distinguishable amongst the different disjunctive types, although a tendency for higher rates of exhaustivity was observed for complex disjunctions. We argued that this tendency, which parallels the significant result obtained with exclusivity inferences, is supported by the view that prosodic prominence (associated with complex disjunctions) cues hearers to interpret utterances on their stronger, exhaustified, parses involving local strengthening of the disjuncts. This tendency for complexity to lead to increased exhaustiveness favours the type of explanation we advanced here, whereby [**Disj** A **Disj** B] puts the focus on the independent disjuncts, biasing towards substitution alternatives $Alt = \{A, B, C, \dots\}$.

Further investigation into the realm of meaning variation among disjunctions is undoubtedly called for. We already have evidence from the domain of existential quantifiers that indefinites come in different epistemic varieties (see, e.g., work by Alonso-Ovalle and Menéndez-Benito 2010, Aloni and Port 2010, Fălăuş 2014, to name only a few). And even from the domain of disjunctions we have evidence that such specialized disjunctions do occur. For example, Ivlieva (2016) notes that some complex Russian disjunctions (*to li… to li* and *ne to… ne to*) give rise to obligatory, non-cancellable ignorance inferences, as shown by the contrast in (14). This is particularly striking because this inference persists even under existential modals, an environment where disjunctions normally give rise to free choice permission inferences and ignorance is obviated.

(14) Ty možeš vzjat' **to li** jabloko, **to li** apel'sin **/ne to** jabloko, **ne to** apel'sin
You may take *to li* an orange *to li* an apple/ *ne to* an orange *ne to* an apple
$\not\rightsquigarrow$ you may take an orange and you may take an apple
$\rightsquigarrow$ it is either an orange or an apple that you're allowed to take

A more detailed investigation is needed but what seems to be the case is that certain inferences are more likely to be lexicalized, with ignorance and free choice being such inferences, to the exclusion of exclusivity inferences (Maria Aloni p.c.). Our results indicate that exhaustive inferences most likely fall in the same category as exclusivity inferences in their resistance to lexicalization. A proper understanding of this pattern will have to be left for another time, but we believe the *Neglect Zero* approach advocated by Aloni (2022) may pave the way towards a solution.

Future work on this topic could also look at the specifics of the morphological makeup of complex disjunctions and what points of variation are observed there. For example, one dimension of variation could relate to the number of morphemes in a given disjunction (e.g. two in *ou bien* versus only one in *soit*). This distinction cuts both within complex disjunctions as well as between simplex and complex, with simplex *i* and complex *ili* being a prime example since *ili* is morphologically made up of the disjunctive marker *i* and the question particle *li*. An even more specific dimension of variation could be formulated in terms of morphological containment, with both *i* vs *ili* and *ou* vs *ou bien* acting as prime examples, since the complex variants are built off of the simplex variants. Suffice it to say, the possible levels of variation are numerous and coming up with any concrete hierarchy of markedness requires significantly more empirical work.

References

Aloni, M. (2022, June). Logic and conversation: The case of free choice. *Semantics and Pragmatics 15*(5), 1–60.

Aloni, M. and A. Port (2010). Epistemic indefinites crosslinguistically. In *Proceedings of NELS 41*.

Alonso-Ovalle, L. and P. Menéndez-Benito (2010). Modal indefinites. *Natural Language Semantics 18*(1), 1–31.

Baker, R., R. Doran, Y. McNabb, M. Larson, and G. Ward (2009). On the non-unified nature of scalar implicature: An empirical investigation. *International Review of Pragmatics 1*(2), 211 – 248.

Bates, D., M. Mächler, B. Bolker, and S. Walker (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software 67*(1), 1–48.

Chierchia, G. (2006). Broaden your views: Implicatures of domain widening and the 'logicality' of language. *Linguistic Inquiry 37*(4), 535–590.

Chierchia, G., D. Fox, and B. Spector (2012). Scalar implicatures as a grammatical phenomenon. In C. Maienborn, P. Portner, and K. von Heusinger (Eds.), *Semantics: An International Handbook of Natural Language Meaning* (Berlin & Boston: de Gruyter ed.), Volume 3, pp. 2297–2332. New York, NY: Mouton de Gruyter.

Degano, M., S. Ramotowska, P. Marty, M. Aloni, R. Breheny, J. Romoli, and Y. Sudo (2023). The ups and downs of ignorance. under review.

Doran, R. B., G. Ward, M. Larson, Y. McNabb, and R. Baker (2012). A novel experimental paradigm for distinguishing between what is said and what is implicated. *Language 88*, 124–154.

Fox, J. and S. Weisberg (2019). *An R Companion to Applied Regression* (Third ed.). Thousand Oaks CA: Sage.

Fălăuş, A. (2014). (Partially) Free Choice of Alternatives. *Linguistics and Philosophy 37*(2), 121–173.

Gazdar, G. (1979). *Pragmatics: Implicature, Presupposition, and Logical Form*. New York, N.Y.: Academic Press.

Gotzner, N., D. Barner, and S. Crain (2020). Disjunction triggers exhaustivity implicatures in 4- to 5-year-olds: Investigating the role of access to alternatives. *Journal of Semantics 37*, 219–245.

Grice, P. (1967). Logic and conversation. Unpublished lecture notes from William James Lectures at Harvard, 1967.

Harrell, F. (2023). *Package 'Hmisc'*. R package version 5.1-0.

Horn, L. (1984). Towards a new taxonomy for pragmatic inference: Q-and R-based implicature. *Meaning, form and use in context*.

Horn, L. R. (1972). *On the semantic properties of logical operators in English*. Ph. D. thesis, University of California Los Angeles.

Ivlieva, N. (2016). Epistemic disjunction and obligatory ignorance. Handout at 'Disjunction days'.

Katzir, R. (2007). Structurally defined alternatives. *Linguistics and Philosophy 30*(6), 669–690.

Kratzer, A. and J. Shimoyama (2002). Indeterminate pronouns: The view from Japanese. In Y. Otso (Ed.), *Tokyo Conference on Psycholinguistics*, Volume 3, pp. 1–25.

Lassiter, D. (2014, 07). Epistemic Comparison, Models of Uncertainty, and the Disjunction

Puzzle. *Journal of Semantics 32*(4), 649–684.
Lassiter, D. (2020). *Graded Modality*, pp. 1–25. John Wiley & Sons, Ltd.
Mandelkern, M. and K. Dorst (2022). Assertion is weak. *Philosophers' Imprint 22*(n/a).
Meyer, M.-C. (2013). *Ignorance and grammar*. Ph. D. thesis, Massachusetts Institute of Technology, Cambridge, MA.
Moss, S. (2015, March). On the semantics and pragmatics of epistemic vocabulary. *Semantics and Pragmatics 8*(5), 1–81.
Nicolae, A. C. (2017). Deriving the positive polarity behavior of plain disjunction. *Semantics and Pragmatics 10*(5), 1–21.
Nicolae, A. C. and U. Sauerland (2016). A contest of strength: *or* versus *either or*. In N. Bade, P. Berezovskaya, and A. Schöller (Eds.), *Sinn und Bedeutung 20*, pp. 551–568.
R Core Team (2023). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
Santorio, P. and J. Romoli (2017). Probability and implicatures: A unified account of the scalar effects of disjunction under modals. *Semantics and Pragmatics 10*(13), 1–61.
Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguistics and Philosophy 27*(3), 367–391.
Spector, B. (2014). Global positive polarity items and obligatory exhaustivity. *Semantics and Pragmatics 7*(11), 1–61.
Swanson, E. (2016). The application of constraint semantics to the language of subjective uncertainty. *Journal of Philosophical Logic 45*(2), 121–146.
Szabolcsi, A. (2015). What do quantifier particles do? *Linguistics and Philosophy 38*(2), 159–204.
van Tiel, B., E. Van Miltenburg, N. Zevakhina, and B. Geurts (2016). Scalar diversity. *Journal of Semantics 33*(1), 137–175.
Yalcin, S. (2007). Epistemic modals. *Mind 116*(464), 983–1026.
Yalcin, S. (2010). Probability operators. *Philosophy Compass 5*(11), 916–937.

Degrees and depiction- gradability in sign languages[1]

Natasha THALLURI — *Harvard University*
Kathryn DAVIDSON — *Harvard University*

Abstract. Based on variation in spoken languages, the Degree Semantics Parameter (DSP) proposes a split between languages that use a degree based system and those that use a delineation system (Beck et al., 2009; Bochnak, 2015). When it comes to sign languages, it has recently been proposed that the phonological form of gradable predicates can iconically represent the boundaries and points on a degree scale, as in Italian Sign Language [LIS] (Aristodemo and Geraci, 2018). From this perspective, sign languages seem to offer visible evidence of abstract linguistic objects like degree scales which have been theoretically motivated in spoken languages but whose existence has been inferred through certain syntactic and semantic cues. However, evidence for a degreeless semantics for American Sign Language [ASL] (Kouli-dobrova et al., 2023) suggests that sign languages could vary as much as spoken languages within this domain.We argue for an alternative semantics for comparative constructions in sign languages with a iconic component in them. Rather than assuming that that sign languages vary with respect to whether this iconicity encodes degrees, we suggest a unified view of all these constructions where the iconicity is analysed as gestures or demonstrations in the sense of Davidson (2015). Under this view, iconicity is insufficient to motivate a degree ontology in a sign language because the linguistic parts of sign languages, being languages, are built around abstraction, and what may appear to be iconic/visible pieces of the grammar are more accurately viewed as gestural depiction, just like spoken language gestures.

Keywords: semantics, sign language, gesture, gradability, degrees, iconicity.

1. Introduction

Formal semantic analyses of natural language phenomena have typically focused on descriptive symbolic meaning. In contrast, recent work on meaning in the visual modality has sparked the need for an updated approach to meaning that can also capture the semantic contribution of depictive or 'iconic' elements that are produced as part of an utterance but are often left unanalysed. Within the domain of spoken languages, this has led to sometimes analogous proposals for the semantic contribution of co-speech gestures in the context of spoken language expressions (Schlenker, 2021; Ebert et al., 2020; Esipova, 2019). The question of how to approach iconic content within formal semantics is particularly salient when modelling natural language phenomena in sign languages where both the depictive and symbolic elements of communication are expressed in the visual modality, where categorizing "gestures" is not a simple matter of modality (even if it were in spoken languages). Although formal semantic work on sign languages in role shift, anaphora, quantification etc. have provided valuable insight, one area that has remained comparatively understudied are gradable expressions in sign languages.

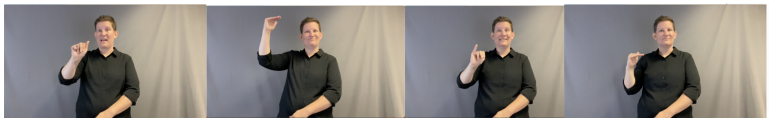
We typically encounter the notion of gradability in the meanings of predicates such as *tall, heavy* or *expensive*. The meaning of a sentence like *the textbook is heavy* seems to be depen-

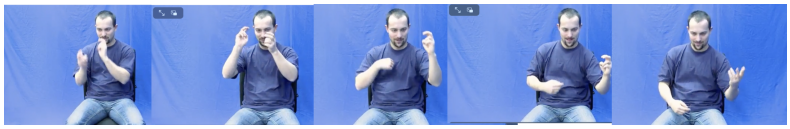
[1]Our most sincere gratitude to Ryan Bochnak for detailed comments on an earlier draft of this paper and very helpful discussion and to Chloe Frey for her assistance in data collection.

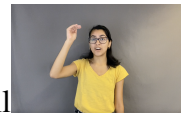
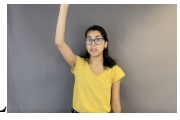
 In: Baumann, Geraldine, Daniel Gutzmann, Jonas Koopman, Kristina Liefke, Agata Renans, and Tatjana Scheffler (eds.) 2024. Proceedings of Sinn und Bedeutung 28. Bochum: Ruhr-University Bochum. 1114-1132.

dent on what exactly classifies as heavy in a particular context. While a weight of 2kgs might be heavy in the context of books, it might not be heavy in the context of furniture. Adding degree variables to our semantic ontology captures this intuition by defining these gradable predicates as a relation between the argument of the predicate and the degree to which the gradable predicate holds of the argument, relative to a contextually determined standard of what classifies as *tall* or *heavy* or *expensive*. Natural language contains a wide range of expressions that make use of the concept of gradability such as comparatives, superlatives, measure phrases, etc. Although well studied in spoken language (Kennedy and McNally, 2005; Kennedy, 2007a), the semantics of gradability in the visual modality is much less well understood. As this paper will argue, sign languages offer a valuable insight into how these concepts of gradability are expressed in the visual modality, and in doing so provide new insight into modeling iconic content in sign languages and gestures. To see why, consider the following sentences in (1)-(4) which are all comparative constructions involving the predicate *tall*. Not only do they all express essentially the same meaning, i.e, *A is taller than B* but they also involve an iconic component that seems to contribute this meaning. The iconic element in question here is the position of the hand in the signing space which depicts the height of the referent and seems to be almost identical in all four cases. In LIS in (1), the position of the hand is represented by the subscripts α, β and γ which refer to vertically ordered loci in the signing space, following the glossing scheme in Aristodemo and Geraci (2018). The DGS example in (3) shows multiple occurrences of the hand-shape TALL2A* high in the signing space to refer to many tall individuals, and the same hand-shape low in the signing space to refer to a short individual.

(1) MAN TALL-$_\alpha$POS$_\beta$ WOMAN TALL- $_\beta$ICONIC-MORE$_\gamma$. IX$_\beta$ 1 METER 70. IX$_\gamma$ 1 METER 80
'Maria is taller than Gianni. This one (Gianni's degree) is 1 meter 70 and that one is 1 meter 80.' (LIS, Aristodemo and Geraci (2018))

(2) 
MARY TALL$_{(at-signer'-head)}$ GIANNI TALL$_{(neutral-space)}$
Lit. 'Mary is this tall, Gianni is this tall.' ('Mary is taller than Gianni.')
(ASL, adaptation from Koulidobrova et al., 2023)

(3) 
TALL2A* TALL2A* I2 SMALL 3 GEST-OFF*
'All the others were taller than me; I'm just too short' (DGS, Konrad et al., 2020)

(4) Alex is tall  but Jo is like TALL 
Interpretation: Both Alex and Jo are tall, but Jo is very tall. (English w co-speech gesture)

This paper aims to answer two questions that arise from the data above. The first concerns how

sign languages encode the semantic concept of gradability and the role they play with respect to the Degree Semantics Parameter. The second is about how sign languages use iconicity, and how to model iconicity within our formal semantic analysis. As seen in (1)-(4), iconic depictions of certain gradable predicates seems to be a common strategy in expressions of gradability in the visual modality, both sign and gesture, and a theory of gradability for sign language should include a proposal for the exact semantic contribution of these iconic expressions. This of course raises the broader question of the semantics of iconic expressions across multiple phenomena in both the visual and spoken modality. We argue for a account that uniformly models the iconicity across all of these gradable expressions as demonstrations.

2. Gradability in natural language

The semantics of gradability remains a relatively understudied phenomenon in sign languages compared to spoken languages. While the conceptual notion of gradability seems to be expressed in every language, the syntax and semantics of expressions such as comparatives, superlatives, measure phrases, degree questions, etc. that encode gradability are subject to a high degree of cross-linguistic variation (Bobaljik, 2012; Bochnak, 2015; Beck et al., 2009). One influential idea in this literature comes from Beck et al. (2009) who propose that the ontology of a degree variable is a point of parametric variation. This is the spelled out as the Degree Semantics Parameter (DSP) in (5).

(5) Degree Semantics Parameter (DSP):
A language does/does not have gradable predicates (type $\langle d, \langle e, t \rangle \rangle$ and related), i.e. lexical items that introduce degree arguments. (Beck et al., 2009)

This proposal challenges the assumption that the concept of a degree is a semantic universal, and that some semantic categories might be subject to parametric variation. The DSP proposes a split between languages that employ a Delineation style semantics (Klein, 1980; Burnett, 2015), and those that use a Degree semantics (Kennedy and McNally, 2005; Kennedy, 2007b; Wellwood, 2015). While the nature of this parameter is an interesting discussion in its own right, for the purposes of this paper, we will assume that the DSP causes languages to pattern differently with respect to whether the ontological category of a degree exists in that language. Note however, that this is not to say that languages without degrees do not express concepts of gradability but rather that they employ a different syntactic and semantic inventory, and thus do so in a different way. We can see this by comparing English, which is a +DSP language according to this proposal, to Washo which has been argued to be a -DSP language (Bochnak, 2015). In the case of comparative expressions, the presence of a lexical item *-er* or its analytic counterpart *more*, as in English, is typically analysed as an operator over degrees, and is often indicative of the presence of a degree ontology in a language. In contrast, the main strategy of comparison in Washo is a conjunction construction which does not have any comparative morphology on the gradable predicate or the standard of comparison (6a). These constructions are not felicitous in crisp judgement contexts i.e cases where there are fine grained differences between two entities with respect to some gradable property. For such cases where the English *-er* comparative is reported to be permitted, Washo uses modifiers like the intensifier šemu ('really') or wewš ('almost') rather than the conjoined comparative as in (6b).

(6) a. *Conjoined comparative*
mé:hu delkáykayiʔ k'éʔi šáwlamhu ʔilkúškuši ʔaš
mé:hu de-ʔil-kaykay-iʔ k'-eʔ-i šawlamhu ʔil-kuškuš-iʔ-aʔ-š
boy NMLZ-ATTR-tall-ATTR 3-COP-IPFV girl ATTR-short-ATTR-AOR-SR
'The boy is taller than the girl'

b. *Comparatives with modifiers*
wí:diʔ behezinaš lák'aʔ wí:diʔ t'í:yeliʔ **wéwši**
wi:diʔ behezin-aʔ-š lak'aʔ wi:diʔ t'i:yeliʔ **wewš-i**
this small-AOR-SR one this big almost-IPFV
this one is bigger than that one
(lit: this one is small, that one is almost big) (Bochnak, 2015)

This is in contrast to a +DSP language like English where the main strategy of comparison is a construction with overt comparative morphology on the gradable predicate which is also the preferred construction in crisp judgement contexts. While English also has expressions for comparison that use a coordination structure (7c) or an intensifier (7b), they seem to have a more restricted usage than the *-er* comparative (7a).

(7) a. John is taller than Mary.
b. John is tall, but Mary is really tall.
c. John is tall but Mary is not.

3. Gradability in sign languages- LIS and ASL

We turn to similar data in the visual modality from two different sign languages that have been argued to be on opposite sides of the DSP, just like English and Washo. As we will see, unlike with spoken languages, these proposals crucially need to capture what seem to be additional iconic components of these gradable predicates. Intriguingly, Aristodemo and Geraci (2018) propose to account for this iconicity in LIS by analyzing the iconic component as a visual representation of the underlying degree component proposed to be covert in spoken languages. However, as Koulidobrova et al. (2023) argue, this analysis seems to make incorrect predictions for ASL, a sign language that has similar iconic expressions of comparison as in LIS, but seems to differ in terms of other expressions of gradability, which point to ASL as falling on the Washo side of the DSP. Crucially, it seems that other expressions of gradability in ASL are best analysed without degrees, in which case a degree-based analysis of iconicity would be incompatible with broader facts of ASL, yet they acknowledge they lack a formal semantic analysis of these iconic comparatives that is compatible with the degree-less semantics suggested for ASL.

3.1. Degrees in LIS

This section reviews the proposal in Aristodemo and Geraci (2018) for visible degrees in LIS, which originated the discussion of degree semantics in sign languages and raised an intruiging case of 'visibility' of logical form. While LIS is essentially supposed to be like other +DSP

languages, the authors highlight two unique aspects of this phenomenon in LIS. The first is that the phonological form of certain gradable predicates seems to encode an 'iconic scale'. This means that the movement of the hand in signing space depicts an abstract scale (vertical or horizontal) with points or loci along that scale representing overt degree arguments. A consequence of this is the second important aspect of gradability in LIS. Overt expression of degree arguments in the language allows for the possibility of anaphoric reference to degrees. As we know, languages with a degree semantics like LIS are fairly common among spoken languages. However, as the authors note, English allows anaphoric reference to individuals but not to the degree itself.

While a degree semantics does not explain why such reference is not possible in all +DSP languages, it does seem to be available in LIS. So one of the pieces of evidence for a degree semantics in LIS is the anaphoric potential of degrees in this sign language. The authors are primarily concerned with open scale gradable adjectives like *tall, deep, expensive, etc*. All of these adjectives employ scales of some dimension of measurement along which a set of degrees is ordered. Their proposal for an Iconic Degree Scale in (8) is a modification of this idea.

(8) Iconic Degree Scale
An iconic scale is the order-preserving mapping of a set of ordered degrees onto a set of ordered points in the signing space (i.e. a line on the horizontal, vertical or lateral plane). Each degree of the scale is represented as a point along a line. (Aristodemo and Geraci, 2018)

Access to this iconic scale however, is subject to certain morpho-phonological constraints. Crucially, the signs for the gradable predicates must be size and shape classifiers, and the movement of the sign must be perpendicular to the plane of articulation. So the LIS sign for DEEP which is semantically the same kind of open scale gradable predicate as TALL, does not have access to this iconic scale. The movement of the sign DEEP is iconic, perpendicular to the plane of articulation with the non-dominant hand moving downward to indicate depth, but it is not a size and shape classifier and so will not express degrees iconically. But for a predicate like TALL, which does satisfy these constraints, the positive form of the predicate, and the comparative form of the predicate are in (1). α,β, and γ refer ordered points on the iconic scale where $\alpha \prec \beta \prec \gamma$. The movement of the hand from one locus to the next is analysed as a morpheme within this analysis. So the movement $\alpha \rightarrow \beta$ is assumed to be the bound *pos* operator and the movement $\beta \rightarrow \gamma$ is the bound morpheme which represents the comparative operator, like the English *-er*. Moreover, in (1), β is the degree to which the man is tall, and γ is the degree to which the woman is tall along the Iconic Degree Scale.

(9) ADRIATIC DEEP AEGEAN MORE
'The Aegean sea is deeper than the Adriatic sea' (Aristodemo and Geraci, 2018)

Other gradable predicates in LIS which do not meet these constraints are not compatible with the bound morpheme. Instead, in these cases the analytic form of the comparative morpheme, glossed as MORE as in (9) is used, and there seems to be no overt morpheme for the *pos* operator in these cases. Moreover, all the adjectives which can use the iconic scale are also compatible with the analytic form. This seems to be indicative of some element of optionality in making use of the iconic scale. Both the analytic MORE and the synthetic $_{\beta}$ICONIC-MORE$_{\gamma}$ have been argued to encode the same semantics of a clausal comparative *-er* operator (10).

(10) a. $[\![-er_{clausal}]\!] = \lambda P_{\langle d,t\rangle}.\lambda Q_{\langle d,t\rangle}.Max(Q) > Max(P)$

However, note that even though the two may be different morphological realisations of the same operator, it is only $_{\beta}$ICONIC-MORE$_{\gamma}$ which also includes overt realisations of degree arguments as loci (locations in signing space) ordered along the iconic scale. This would mean that gradable predicates which do not have access to the iconic scale presumably do not allow anaphoric reference to degrees since it would not be possible to assign them to visible loci. In sum, the morphology that makes overt reference to degrees is restricted to a certain class of predicates, and is always optional. In particular, it seems to depend in part on how 'iconic' a particular sign is, with iconicity in this context being defined by certain morphophonological constraints i.e, they must be classifiers with a specific direction of hand movement.

Aristodemo and Geraci (2018) argue for a degree ontology in LIS on the basis of the fact that certain gradable predicates in this language allow an overt realisation of degree arguments. Perhaps not surprisingly, the existing typology of comparatives in spoken languages in Beck et al. (2009) does not discuss overt realisation of degree arguments as a diagnostic for a degree ontology in a language, so the argument is based on entirely disjoint diagnostics. Moreover, the possibility of anaphoric reference to degrees in LIS discussed by Aristodemo and Geraci (2018) is a novel observation since as the authors note, this sort of anaphoric reference has not been observed in +DSP spoken languages like English, but it remains unclear whether this sort of reference to degrees is not possible in English by virtue of its modality, or whether there are other factors at play. If LIS as a +DSP language allows anaphoric reference to degrees, then it would certainly be interesting to understand whether other + DSP languages pattern similar to LIS, and particularly whether other sign languages also allow this property. As the facts stand, it would seem that any sign language where gradable predicates can make use of the iconic degree scale should allow anaphoric reference.

This brings us to a broader question about the theory of iconicity being assumed for LIS. It is very likely that gradable predicates whose morphophonological form satisfies the constraints outlined in Aristodemo and Geraci (2018) can be found across sign languages. Does this proposal then predict the possibility of visible degrees in all those cases in different sign languages? If so, one would assume that all sign languages employ a degree semantics. However, as we will in see in the next section, Koulidobrova et al. (2023) argue that ASL differs from LIS in this regard, and may be best analyzing using a delineation or degree-less semantics instead.

3.2. A degree-less semantics in ASL

Contrary to previous work on gradability in ASL (Wilbur et al., 2012; Kentner, 2020) which assumes the presence of degrees in ASL, Koulidobrova et al. (2023) argue that a degree-less analysis along the lines of Washo better explains the data from ASL. According to their findings, differential measure comparatives across a wide range of gradable predicates are strictly ungrammatical in ASL (11a). It is also not possible to construct a measure phrase with an overt degree in the attributive position as in (11b).

(11) a. *ALEX a-IX TALL$_{(neutral-space)}$ JO b-IX TALL$_{(at-signer'-head)}$ 4 INCH
'Jo is 4 inches taller than Alex'

b. *BOOK (IX) 4 KILO HEAVY NOT 1-POSS
That 4 kilo heavy book- not mine (Koulidobrova et al., 2023)

If this is right, then according to the diagnostics in Beck et al. (2009), ASL does not seem to have the kind of expressions that refer to degrees. When it comes to expressing comparison, there are multiple strategies in ASL; a juxtaposition construction like in Washo without an overt comparative marker; comparison using a non-manual intensification marker; an overt lexical item such as MORE, BEAT, BETTER, SAME,etc. A further strategy which the authors note as being the most common and intuitive one is comparative depiction, shown (2). This expression has exactly the same components as the explicit comparative in LIS, with the same gradable predicate TALL which seems to have the same form as its LIS counterpart.

The argumentation for ASL essentially goes in the opposite direction to what has been proposed for LIS, which leads us to radically different analyses for a construction which in fact looks very similar in the two languages. Koulidobrova et al. (2023) argue that ASL systematically lacks constructions one might expect to see in a +DSP language and hence it reasonable to propose that it is a -DSP language, so the sentence in (2) cannot contain degrees. Meanwhile, the proposal for LIS argues that similar constructions like (1) contain overt reference to degrees, and hence provide a unique form of evidence for a degree ontology in LIS.

Clearly, we need a resolution. We agree with Koulidobrova et al. (2023) that a degree-less story for ASL is appealing in light of several empirical patterns in the language, but in order for the degree-less analysis of ASL to stand, there needs to be a formal semantic analysis that explains the constructions in (12) without reference to degrees. Koulidobrova et al. (2023) allude to two possible directions for these cases. For (12a), they suggest that this could be a case of comparative depiction or demonstration in the sense of Davidson (2015) rather than visible degrees. And in (12b), the intensification of the sign FAST is a kind of predicate modifier which results in an intensification interpretation like the modifier *šemu* in Washo (Beltrama and Bochnak, 2015), running contrary to the claim in Wilbur et al. (2012) which analyses intensification in ASL as a form degree modification.

(12) a. a-IX a-HEAVY b-IX HEAVY (*minimal downward movement*)
'B is a little heavier than A'
b. a-IX a-FAST b-IX $\overline{\text{FAST}}^{intens}$
' A is fast, B is a little faster' (Koulidobrova et al., 2023)

Note that both expressions involve some modification of the phonological form of the sign, although the intensification strategy is not strictly restricted to 'iconic' signs. The existing literature on constructions of this form has assumed a degree-based analysis, crucially relying on the iconicity as evidence for degrees in these constructions. Both in the extension of the Event Visibility Hypothesis in Wilbur et al. (2012), as well as the Iconic Degree Scale in Aristodemo and Geraci (2018), the claim has been that the phonological form expresses the semantic scales that are a part of the lexical semantics of these gradable predicates.

This paper will put aside the question of intensification, which is best examined in the context of intensification cross-linguistically. In the next section we focus on the non-intensification examples, and argue in agreement with Koulidobrova et al. (2023) that cases such as (12a) are best suited to a demonstration based account. Taking on the challenge they note for seman-

tics, we propose an degree-less analysis for comparative depiction in ASL where the iconic component of the sign for the gradable predicate is analysed as a demonstration, providing a degree-less analysis for this data. Moreover, given the similarity between the constructions in LIS and ASL in (3) and (2), we also argue that such an analysis significantly weakens the claim in Aristodemo and Geraci (2018) for a degree ontology in LIS which is based solely on the iconicity in these constructions

4. Iconicity via demonstration

4.1. The semantics of demonstrations

The original argument for the concept of a demonstration in Clark and Gerrig (1990) takes on the question of the 'iconicity' in quotations, namely the fact that a quote has to report a speech event in roughly (although not always exactly) the form that it was originally made, and thus analyses quotations (in written and spoken language) as a performance or a demonstration. Davidson (2015) builds on this idea to propose a compositional analysis for demonstration in language in quotations and in other iconic gestural phenomena, both spoken and signed. Under this view, in order to compose a demonstration d_v (a communicative event of type v) with a natural language expression, an operator establishes a relation between a demonstration and the event denoted by the expression. It could be a covert operator, or it could be denoted by a lexical item as in the English '*like*' quotations in (13d). Existing implementations of this approach to iconicity have been proposed for spoken language quotations and sign language classifiers (Davidson, 2015; Zucchi, 2017), as well as role shift (Davidson, 2015; Maier, 2018; Steinbach, 2023).

(13) a. A demonstration d is a demonstration of e (i.e. demonstration(d,e) holds) if d reproduces properties of e and those properties are relevant in the context of speech.

b. $[\![demonstration]\!] = \lambda d.\lambda e[demonstration(d,e)]$

c. $d_1 =$ Chloe's utterance "that's a huuuge dog"

d. [[Chloe was like "that's a huuuge dog"]] =
$\exists e[Agent(e,Chloe) \wedge demonstration(d_1,e)]$ based on (Davidson, 2015)

Sign language classifier predicates can be decomposed into a linguistic component (contributed by the handshape), and a gestural component which is expressed with iconic uses of the classifier sign's location and movement. The demonstration tool in (13b) allows us to model the semantic contribution of the non-linguistic component of the classifier. In an event semantics framework, the classifier takes as its arguments a demonstration and an event (this can be an agent and/or a theme depending on the type of classifier). This is illustrated in the denotation of the size and shape classifier for BOOK in ASL from Davidson (2015).

(14) $[\![\text{CL-B MOVE}]\!] =$
$\lambda d.\lambda x.\lambda e.[theme(e,x) \wedge flat-object(x) \wedge moving(e) \wedge demonstration(d,e)]$

The resulting expression has an interpretation along the lines of *'the book moved in a manner that resembles demonstration d'*. The next section illustrates how this analysis of size and shape

classifiers in ASL can be straightforwardly extended to iconic gradable predicates in ASL.

4.2. Comparative depiction as demonstration

The examples of comparative depiction in Koulidobrova et al. (2023) all involve adjectives which have a strong iconic component in the form of the sign.[2] For these adjectives we propose a similar split between the linguistic component and the demonstration component of the sign. The linguistic component in this case is a gradable predicate whose denotation is determined by the contextually specified comparison class which is encoded by the handshape, i.e. the linguistic aspect of the sign TALL can simply mean roughly 'tall for the context'. The demonstration component is expressed by the movement and location of the hand in the signing space, exactly as in the case of the classifiers. In this way, we assume that iconic movements are integrated into the form of the sign as demonstrations of type δ. So when we think of a sentence like ALEX TALL(signed in neutral space) *'Alex is tall'* in ASL, the location of the hand in the signing space is the demonstration. Just as in the case of the classifiers, the demonstration composes with TALL by means of a covert *demonstration* predicate.

In order to capture the semantics of gradable predicates within this framework, we assume that gradable adjectives represent states, not events. This is in some sense a natural extension of Davidson (2015), which focuses on manner classifiers which are eventive. Since sign languages also have classifiers (e.g. size and shape predicates) that denote states, this approach can be extended to a broader range of classifiers not just those that are relevant to gradability. Consider the spoken language example in (15) where the demonstration event depicts the state of exhaustion that the person is in.

(15) $[\![$ Chloe was like exhausted$_{\textit{drooping shoulders}}]\!] =$
$[\exists s.(Exhausted(s) \wedge Holder(s, Chloe) \wedge Demonstration(d, s))]$
(Interpretation: Chloe was really tired and the demonstration event depicts her physical state.)

We now have the first component of our analysis, which is an analysis of demonstrations that express states. Next, we turn to other pieces of formal machinery that are required to make the semantics of gradable predicates compatible with the account of iconicity above. Within an event semantics framework, gradable predicates can be predicates of events or states, and an additional functional head v_S introduces the thematic role of the Holder of the state (16f). Since we are primarily concerned with gradable adjectives in this paper, we adopt a semantics of gradable adjectives from Cariani et al. (2023). Within this approach, gradable adjectives like TALL and HOT do not encode a relation between degrees and individuals in their lexical semantics. Instead, they are defined as properties of states. A contextually determined comparison class (Klein, 1980) is encoded here as a background ordering of states which is part of the the lexical semantics of the gradable adjective. Thus, TALL is a predicate of states, with a contextually determined threshold property of being tall that maps a state s to true if the state meets the relevant threshold in context c and if the state is a relevant state in the domain of

[2]Koulidobrova et al. (2023) note that all the gradable predicates they tested which employ a comparative depiction strategy are also compatible with a different strategy of comparison which involves the use of MORE or BEAT, similar to the optionality between the synthetic and analytic forms in LIS

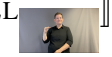
height ordering (16a).

(16) a. $[\![tall]\!] = \lambda s : s \in domain(D_{height} \succeq_{height}).tall_c(s)$ [based on Cariani et al. (2023)]

b. $[\![demonstration]\!] = \lambda d.\lambda s, demonstration(d,s)$

c. [] $= \delta_1$

d. $[\![demonstration([\![$  $]\!])]\!] = \lambda s.demonstration(\delta_1, s)$

e. $[\![$ TALL  $]\!] = \lambda s(s \in domain(\langle D_{height}, \succeq_{height} \rangle)).Tall_c(s)$
$\wedge Demonstration(\delta_1, s))]$

f. $[\![V_S]\!] = \lambda x.\lambda s.Holder(s)(x)$

g. $[\![$ ALEXa-IX TALL  $]\!]^c$
$= [\exists s(s \in domain(D_{height}, \succeq_{height})).Tall_c(s) \wedge (Holder(s, Alex)$
$\wedge Demonstration(\delta_1, s))]$
'Alex is in a state of being tall (relative to the context c) and δ_1 demonstrates that state.'

The implicit *demonstration* predicate (16b) first takes as an argument the iconic component of the sign δ_1, which is the position of the hand in neutral space and returns a predicate of states. It then combines with the gradable adjective and the holder of the state via Predicate Modification to give us the truth conditions in (16g). This is the positive form of the gradable adjective, produced in a way to iconically express Alex's height.

For the comparative form of the gradable adjective in a language like English, Cariani et al. (2023) introduce degrees through the denotation of *-er* which introduces a measure function and combines with any adjective phrase, NP or VP if they are measurable. It imposes a condition on the background ordering of states and requires that the measure function maps s to a degree greater than any denoted by the *than*-clause. However, since we are assuming a degree-less semantics for ASL, there is no comparative operator like *-er* in the ASL cases of comparative depiction. For the example in (2) we simply see two instances of the positive form of the gradable adjective combining with two different demonstration arguments δ_1 and δ_2. Combining all these elements, we can the following truth conditions for a comparative depiction in ASL (simplifying the information structural features of having the proper name followed by a pronoun).

(17) a. ALEX a-IX TALL  JO b-IX TALL 
Lit. 'Alex is this tall, Jo is this tall.' ('Jo is taller than Alex.')

b. $[\![$ ALEXa-IX TALL  $]\!]^c [\![$ JOEb-IX TALL  $]\!]^c$
$= [\exists s(s \in domain(\langle D_{height}, \succeq_{height} \rangle))(Holder(s, Alex) \wedge Tall_c(s)$
$\wedge Demonstration(\delta_1, s))][\exists s(s \in domain(\langle D_{height}, \succeq_{height} \rangle))(Holder(s, Joe)$
$\wedge Tall_c(s) \wedge Demonstration(\delta_2, s))]$
'Alex is in a state of being tall (relative to the context c) and δ_1 demonstrates that state. 'Joe is in a state of being tall (relative to the context c) and δ_2 demonstrates that state.'

Crucially, in this analysis the demonstration is *not a degree*. Rather, gradable adjectives denote properties of states: there is a state of being tall that is true of an individual, and the demonstration is a reproduction of that state or property. As we can see from the example above, we infer that Joe is taller than Alex since δ_2 demonstrates a greater height than δ_1. Koulidobrova et al. (2023) argue that these constructions are felicitous in crisp judgement contexts (e.g. where they are both tall but one is minimally taller than the other), so the demonstration argument can be used to express fine grained distinctions in height.

These comparative depiction cases are distinct from the juxtaposition constructions in ASL which, like the Washo juxtaposition constructions as in (18b) are only felictious in contexts where the predicate is true of one entity and not of the other. It is predicted that a sentence like (18a) is only possible in contexts where Mary's hair is curly and Paul's hair is straight, but not when both Mary and Paul have curly hair.

(18) a. HAIR left hand-a-MARY-CURLY right hand-PAUL STRAIGHT
'Mary's hair is curly, Paul's hair is straight.' ASL,(Koulidobrova et al., 2023)

b. mé:hu delkáykayi? k'é?i šáwlamhu ?ilkúškuši?aš
mé:hu de-?il-kaykay-i? k'-e?-i šawlamhu ?il-kuškuš-i?-a?-š
boy NMLZ-ATTR-tall-ATTR 3-COP-IPFV girl ATTR-short-ATTR-AOR-SR
'The boy is taller than the girl' Washo, (Bochnak, 2015)

By including the ontology of a demonstration in our semantics; an account of gradability in ASL without degrees becomes possible. As discussed earlier, demonstrations are not modality specific. They can be quotations in spoken or sign language, but they can also be iconic elements of classifiers and gradable predicates. It is also possible that there are predicates that do not have as close an integration with demonstration as the cases discussed above. A gradable predicate like SMART in ASL might not have the same sort of iconicity as predicates like TALL. In such a case, ASL presumably employs a different strategy of comparison like a BEAT construction or intensification. Now that we have seen how demonstrations integrate with the lexical semantics of signs, next we revisit the case of gradability in LIS.

4.3. A reanalysis of LIS

On one hand, assuming an overt comparative operator makes it fairly straightforward to model LIS as a language with a standard degree semantics. However, as we have seen in this section, LIS is different from other spoken languages with a degree semantics in a number of ways. Not only does the iconicity seem to offer evidence for an overt *pos* morpheme, as proposed by Aristodemo and Geraci (2018), but the iconic movement in the signing space also motivates the presence of an overt measure function. This also seems to allow another unusual property of LIS which is the possibility of anaphoric reference to degrees. It is also worth noting that in a typical clausal comparative like (19a), the comparative operator does not take overt degree arguments but involves abstraction over degrees, and under this view the LIS *-er* operator has the same denotation as the English clausal comparative operator. However, the LIS comparative constructions seem to involve a direct comparison between two degree arguments δ_β and δ_γ.

(19) a. #John is taller than Mary. It is 6ft tall.

b. John is taller than Mary. He is 6ft tall. (Aristodemo and Geraci, 2018)

Such an analysis of LIS as a +DSP language seems to involve overt evidence of several components of grammar that had previously only been analyzed as covert in spoken language. Moreover, the presence of these components is very clearly linked to the iconicity in this languages. One reason for this difference could be that we do not see overt expressions of these elements in other spoken languages because they lack they same level of iconic potential that LIS has. As Aristodemo and Geraci (2018) note, the data in LIS supports a degree semantics *because* of the overwhelming evidence of overt degrees in the language. In their view, the degrees exist because we can 'see' them. An analysis using overt degrees provides the truth conditions in (20) for the LIS comparatives. If gradable adjectives themselves do not encode a relation between degrees and individual, but rather a predicate of ordered states as in Cariani et al. (2023), the iconic *pos* in LIS would encode a measure function $A(\mu)$ that introduce degree arguments ordered along a contextually relevant scale.

(20) $[\![\text{MAN TALL }_{\alpha}\text{POS}_{\beta}\text{WOMAN TALL}_{\beta}\text{ICONIC-MORE}_{\gamma}]\!] = (\exists s[Holder(s)(Man) \wedge Tall(s) \wedge A(\mu)(s)(= \delta_{\beta}) \succ \delta_{\alpha}])(\exists s[Holder(s)(Woman) \wedge A(\mu)(s)(= \delta_{\gamma}) \succ \delta_{\beta}])$
'The woman is in a state such that her degree δ_{γ}, the man is in a state such that his degree of tallness δ_{β} and it is the case that δ_{γ} exceeds δ_{β}

Since spoken languages haven't been argued to use overt degrees, a standard degree semantics for comparatives does not include argument slots for overt degree variables; the positive and comparative operators in (20) are a modification of the proposal in Cariani et al. (2023) in order to include the overt degree arguments. The truth conditions in (20) state that the height of the man is of a degree δ_{β} which exceeds the standard degree δ_{α}, and the height of the woman is of a degree δ_{γ} which exceeds the standard degree δ_{α}. In this framework, (20) differs from the English clausal comparative since it explicitly encodes a relation between two over degree variables rather than a comparison relation between two sets of degrees.

5. Extending the DSP to sign languages

5.1. Situating LIS and ASL in the typology

On the face of it, we have two new languages to add to our typology. According to Aristodemo and Geraci (2018), LIS has been analyzed as a +DSP language like English, with some constructions that seem to make use of degree arguments, and apparent operators such as the comparative *-er* operator that manipulates degree arguments. On the other hand, Koulidobrova et al. (2023) argue that, ASL patterns empirically like Washo with respect to several tests for a degree ontology. Constructions such as differential measure phrase comparatives and degree questions which target the degree argument slot are not available in ASL, and so we provided a delineation semantics that incorporates demonstrations to model their iconic component. However, perhaps these languages are themselves not so empirically different. Here, we consider more closely the case of the comparative constructions in crisp judgement contexts. Recall that English and Washo pattern in opposite ways with respect to the DSP, and have vastly different constructions to express comparison between two individuals in such cases. English employs the familiar comparative operator lexicalised as *-er* or *more* which establishes a greater than relation between two predicates of degrees. Washo, which lacks such comparative operators

uses a conjuctive strategy along with modifiers such as *šemu* (really) or *wewš*(almost).

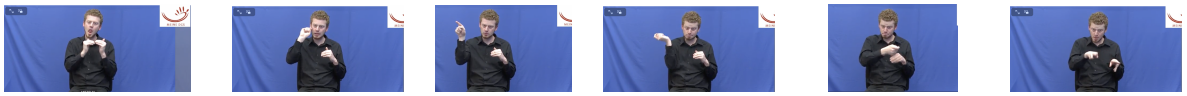
When we examine the constructions used by LIS and ASL to express comparison in crisp judgement contexts, one would expect to see a similar difference between the two languages given the existing claims about the DSP. But strikingly both sign languages seem to make use of essentially the same strategy of comparison. This is completely at odds with the predictions of the DSP. The presence of a degree ontology in a language is supposed to result in significant differences in the strategies used to express comparison. So how can a +DSP and a -DSP language employ what seem to be identical constructions? Taking the existing analyses of these languages on their face, this conflict makes it difficult to predict how other sign languages will potentially fit into the gradability typology. What can we say about a sign language that uses the same strategy of comparison that seems to be available in both LIS (1) and ASL (2)?

To consider this problem in more detail, we turn next to a third sign language which remains unstudied in this domain. The third data point from our initial puzzle is an example of a comparative construction in German Sign Language (DGS) from the DGS corpus (Konrad et al., 2020). Given the similarities (3) has to the LIS and ASL examples, if those two languages really are different then a question arises whether to analyze it as a clausal comparative or a comparative depiction construction; alternatively, if sign languages generally show the same kind of comparative, this may push us further to use a unified story for all.

5.2. Gradability in DGS: evidence from corpus data

Beyond the iconic comparative in DGS[3] in (3), we see a lot of constructions that express comparison similar to ASL as reported by Koulidobrova et al. (2023), including BEAT comparatives and intensification. Where does this put DGS in terms of the DSP?

(21) BEAT comparative



COMPARISON BERLIN1C* THERE1* BERLIN1C* TO-BEAT-7* PARIS
"But Berlin is bigger than Paris in comparison.

(22) Intensifier comparative

LOUD1C OUTDOORS2* $INDEX1
‘(‘Speak louder!’) the people outside were supposed to hear me say it ’ Konrad et al. (2020)

As Koulidobrova et al. (2023) argue, other comparison strategies such as intensifier comparatives as well as BEAT comparatives need not encode a degree semantics in ASL. From that perspective, it seems like DGS might pattern like ASL with respect to the DSP.[4] Of course, the case for a degree semantics (or lack thereof) in DGS must be based on a systematic investigtion

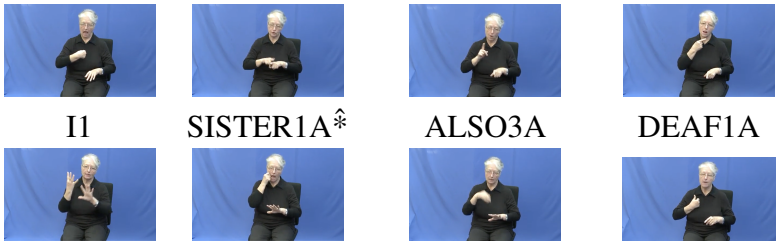
[3]The DGS data presented here follows the glosses in the Hamburg DGS corpus (Konrad et al., 2020)
[4]For a degree-less anaylsis of intensification see Beltrama and Bochnak (2015).

into an entire family of expressions that is indicative of a degree ontology in a language. (23) lists diagnostics for the DSP in a language.

(23)
- •Differential degree comparative.
  A is 5 cm taller than B.
- •Degree comparative
  A is taller than 5 cm
- •Superlative
  A is the tallest person in the room
- •Measure phrase construction
  A is 10 cm tall.
- •Equative
  A is as tall as B is
- •Degree question
  How tall is A? (Beck et al., 2009)

That said, there is some evidence that there might be an English style differential comparatives in DGS, such as (24). This finding was somewhat serendipitous; further empirical corpus and/or elicitation work may more clearly help determine the role of differential degree comparatives, degree comparatives, measure phrases, etc.

(24)



'my sister is deaf as well and she's nine years older than me'

Although a complete analysis of DGS is beyond the scope of this paper, we hope that this comparison of the three languages highlights the tension in existing proposals for gradability in sign languages. It is clear that the choice between the two analyses may not be entirely informed by existing tests, but also depends on whether one assumes that iconicity represents a linguistic concept like a degree, or something extra-linguistic like a demonstration. On the whole, the ASL proposal leans towards a degree-less semantics, while the LIS proposal leans towards a degree semantics, and focusing primarily on these iconic comparatives cannot settle the debate. This is also in line with what we know about the predictions of DSP in spoken languages, where the comparative construction might be indicative of a degree ontology in a given language, but should be considered in light of the rest of the relevant degree constructions. Moreover, the comparative constructions in these three sign languages show a clear uniformity in their use of iconicity that neither the prior proposal for LIS nor the one for ASL captures; their iconicity seems to be coincidental if both are taken at face value. Without a theory that explains the very salient commonality in these cases, it is difficult to investigate the potential-crosslinguistic variation in this domain among sign languages.

We suggest, in an effort towards taking the iconicity to be from a common source, that this difference between the two proposals regarding the comparatives in (1) and (2) is not because LIS has degrees while ASL does not, but rather it is a result of modelling the iconicity in these constructions in two different ways. In the next section, we will argue for a cohesive theory of iconicity that can be generalised over multiple phenomena including co-speech gesture. We follow an extension of Davidson (2015) where iconic elements of language are classified as

demonstrations, a distinct ontological category. This clearly draws distinctions between iconic and abstract elements of language while still allowing both to be integrated into the logical form of the utterance.

6. Towards a unified theory of iconicity

While (Koulidobrova et al., 2023) argue for a degree-less semantics for ASL, this stance is complicated by the fact that BEAT, *intes*, and iconic comparatives have been used to argue for a degree analysis of ASL (Wilbur et al., 2018; Kentner, 2020). However, the absence of constructions like differential degree comparatives and degree questions strengthen the case for a degree-less semantics. Moreover, the demonstration based account for comparative depiction constructions involving gradable predicates completes the picture of ASL as a degree-less language. Analysing iconicity as demonstration also weakens the claims of a degree semantics in LIS which is argued *on the basis of the iconicity* in such comparative expressions.

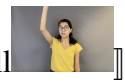
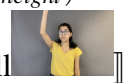
In order to confirm a degree semantics for LIS, we'd want to look for further evidence independent of iconicity in the form of other degree constructions. As it is, the demonstration account is equally compatible with the LIS data which has essentially the same elements as the ASL case. In fact, the demonstration account is further supported by the fact that iconic gradable predicates in LIS *must* be size and shape classifiers, which makes them perfectly compatible with the semantics of classifiers proposed in Davidson (2015), as well as the extension of the proposal to gradable predicates which was laid out in the previous section. In contrast, a degree-based analysis seems potentially more problematic in ASL where there is less evidence for degrees in non-iconic expressions of gradability.

The demonstration based account can also be clearly extended to the DGS example in (3) . The sign for tall (token label TALL2A*), has the same form as the sign for adult (token label (ADULTS1A)). Just as in the ASL case, there is nothing (3) that strictly requires a degree semantics (given our analysis of its iconicity), but like ASL, the argument in favor of a degree-less semantics is merely based on the absence of constructions we expect to find if a language makes use of degrees in this way. Not finding this evidence, either in ASL or in DGS, could be due either to cross-linguistic parametrization, or to the data not happening to contain one (especially in the case of corpus work). Certainly, though, a demonstration account takes away the iconic examples as a focus of this debate, and is a promising direction towards a unified analysis of all such expressions of comparison across sign languages which involve iconic gradable predicates, however they fall on this particular semantic parameter.

The final part of this puzzle is the English comparative accompanied by a co-speech gesture in (4). English is classified as a +DSP language which allows clausal comparatives of the form *A is taller than B is*. The *-er* morpheme is typically analysed as a comparative operator that encodes a greater than relation between two sets of degrees. This is also the analysis initially proposed by Aristodemo and Geraci (2018) for comparatives in LIS. However, the construction in (4) is quite different from the standard clausal comparative. It uses a conjunctive strategy of comparison resembling the cases in -DSP languages like Washo. Rather than a degree operator, we see two independent clauses connected by the conjunction *but* and the gradable predicate in each conjunct is accompanied by a co-speech gesture.

Once again, this construction is remarkably similar to the comparative constructions in ASL and LIS that we have seen so far. Just as in the earlier cases, the co-speech gesture is analysed as a demonstration that combines with the English gradable predicate *tall*. Unlike in the sign language expressions, this case also has the overt *demonstration* predicate which is denoted by the English *like*. If we assume a delineation style semantics, the truth conditions in (25) are identical to the comparative depiction case in ASL that was discussed earlier .

(25) a. $[\![tall]\!] = \lambda s : s \in domain(D_{height} \succeq_{height}).tall_c(s)$

b. [Alex is tall ]^c[Jo is like tall ]^c
$= [\exists s(s \in domain(\langle D_{height}, \succeq_{height}\rangle))(Holder(s,Alex) \wedge Tall_c(s)$
$\wedge Demonstration(\delta_1,s))][\exists s(s \in domain(\langle D_{height}, \succeq_{height}\rangle))(Holder(s,Joe)$
$\wedge Tall_c(s) \wedge Demonstration(\delta_2,s))]$
'Alex is in a state of being tall (relative to the context c) and δ_1 demonstrates that state. 'Joe is in a state of being tall (relative to the context c) and δ_2 demonstrates that state."

If we follow Aristodemo and Geraci (2018) in assuming that the co-speech gestures are overt representations of degree variables, we may be tempted to implement a degree semantics with the co-speech gesture combining with the positive form of the gradable predicate in each conjunct, and the gesture encoding the iconic degree scale. However, it is important to note that (4) is truth conditionally distinct from the explicit comparative without the co-speech gesture in English.

Within a standard Montague semantics, the *-er* operator typically denotes a relation between two sets of degrees. Within an event semantics framework, this denotation must be modified, since gradable predicates denote a predicate of states rather than a relation between degrees and individuals. Instead, an *-er* operator in Wellwood (2015) assumes that the *-er* takes a measure function g, a degree argument d (contributed by the comparative clause), and a state or event of type α (26c). In this case, an explicit clausal comparative without a co-speech gesture would have the truth conditions in (26d).

(26) a. Jo is taller than Alex is
b. $[\![-er_{clausal}]\!] = \lambda P_{\langle d,t\rangle}.\lambda Q_{\langle d,t\rangle}.Max(Q) > Max(P)$
c. $[\![-er_{event}]\!] = \lambda g.\lambda d.\lambda \alpha.g(\alpha) \succ d$
d. True iff $\exists s[Holder(s)(Joe) \wedge Tall(s) \wedge A(\mu)(s) \succ$
$max(\lambda d.\exists s'[Holder(s')(Alex) \wedge Tall(s') \wedge A(\mu)(s') \succeq d])]$
'The degree of the state of tallness of Joe exceeds the degree of tallness of Alex'

Crisp judgement contexts are a crucial diagnostic for explicit comparatives i.e cases where there are fine grained distinctions between two entities with respect to the property of the gradable predicate (Kennedy, 2007a). Consider a case where Joe is 6ft tall and Alex is 5'11 ft tall. The English clausal comparative is felicitous in this context, while (4) is predicted to be degraded or infelicitous here. (26d) is acceptable in a context where both Jo and Alex are tall but there is a minimal difference in their heights. It is also felicitous in a context where both Jo and Alex are short, but Jo is still slightly taller than Alex.

However, comparatives like (4) are not possible in crisp judgement contexts. The assertion in

the implicit comparative is that relative to a context c, the gradable predicate is true of one individual and false of the other. So we can imagine a context where Jo is 7ft tall and ALex is still 5'8 ft tall. Now the truth conditions in (4) entail that in this context, Jo is tall while Alex is not. Moreover, it also entails that both individuals are tall. It is infelicitous in a context where one of them is tall and the other one is short, or in a context where both of them are short.

We predict that all of the instances with an iconic demonstration of height in LIS, ASL, DGS, and English are more like implicit comparatives accompanied by a demonstration rather than explicit comparatives. A further complicating factor in the sign language data is that the signs for the gradable adjectives TALL and SHORT are identical except for the location of the hand in the signing space. While the hand moves up from the neutral space for TALL, it moves down from the neutral space for SHORT. So it is not clear whether the demonstration is combining with the gradable predicate *tall* or whether this is a pure demonstration of height. Crucially, these are all cases of an indirect form of comparison. In the cases like the English clausal comparative, the meaning *A is taller than B* is encoded in the truth conditions of the sentence. However, with the implicit comparatives discussed here, the meaning *A is taller than B* is an inference from the demonstrations. The addressee infers that Jo is taller than Alex due to the fact that the demonstration δ_2 (mapped to Joe) shows a height that is greater than δ_1 (mapped to Alex).

7. Conclusion

This paper presents a demonstration-based account for comparative depiction in ASL which strengthens the case for a degree-less semantics in this sign language. The iconic component of gradable predicates such as TALL can be understood as demonstrations which compose with the predicate via a *demonstration* operator. This is a departure from the analysis proposed for very similar expressions in LIS where it has been claimed that the iconicity in these constructions are components of the logical form such as degrees and scales, opening up room for investigation for the best analyses of each of these languages, and the empirical bases for these choices. We have also shown how this approach can account for comparatives in LIS, DGS, as well as conjoined comparatives in English with co-speech gestures. Analysing iconic elements as demonstrations as shown in this paper allows for a more generalised theory of iconicity that cuts across languages and modalities. We also know from the existing literature that this approach can be extended to other sign language phenomena like role shift and classifiers.

We began with two sign languages that seem to pattern on opposite sides of the Degree Semantics Parameter. Whether this is truly the case remains a question for further research; one would certainly expect to find a similar manner of cross-linguistic variation in sign languages as we find in spoken languages. It may well be the case that ASL patterns like Washo, while LIS and DGS pattern like English and other +DSP languages. However, we make the case that these highly iconic comparative constructions are insufficient to make that claim. A case for degrees in these languages would involve an investigation into classes of expressions, with a wider range of gradable predicates do not encode a demonstration in their form in the manner that TALL does. This would reveal more about the underlying architecture of these constructions in each of these sign languages. More specifically within the class of expressions of comparison,

there remains an intriguing question of the finer grained distinctions between BEAT, and intensification and the depictive strategy discussed here, as well as their cross-linguistic distribution among other sign languages.

The proposal for modelling iconicity presented here does not predict that all sign languages are fundamentally the same when it comes to expressing comparison of gradable concepts. Nor does it argue that iconicity does not have a semantic contribution in these cases. It aims to present a unified approach to iconicity which captures the distinction between the depictive and descriptive elements of utterances. While demonstrations do contribute semantic content via the predicate created by combining the demonstration operator and the language-external demonstration, they are also distinct in several ways from elements that combine compositionally in the abstract linguistic structure of these expressions. We argue that the visual modality does not include the interpretation of iconic components in its core grammar. Instead, demonstrations as a separate module of meaning can, and often do co-occur with linguistic structure.

References

Aristodemo, V. and C. Geraci (2018). Visible degrees in Italian sign language. *Natural Language & Linguistic Theory 36*(3), 685–699.

Beck, S., S. Krasikova, D. Fleischer, R. Gergel, S. Hofstetter, C. Savelsberg, J. Vanderelst, and E. Villalta (2009). Crosslinguistic variation in comparison constructions. *Linguistic Variation Yearbook 9*(1), 1–66.

Beltrama, A. and M. R. Bochnak (2015). Intensification without degrees cross-linguistically. *Natural Language & Linguistic Theory 33*, 843–879.

Bobaljik, J. D. (2012). *Universals in Comparative Morphology: Suppletion, superlatives, and the structure of words*. MIT Press.

Bochnak, M. R. (2015). The degree semantics parameter and cross-linguistic variation. *Semantics and Pragmatics 8*, 6–1.

Burnett, H. (2015). Comparison across domains in delineation semantics. *Journal of Logic, Language and Information 24*(3), 233–265.

Cariani, F., P. Santorio, and A. Wellwood (2023). Positive gradable adjective ascriptions without positive morphemes. In *Proceedings of Sinn und Bedeutung*, Volume 27, pp. 96–113.

Clark, H. H. and R. J. Gerrig (1990). Quotations as demonstrations. *Language*, 764–805.

Davidson, K. (2015). Quotation, demonstration, and iconicity. *Linguistics and Philosophy 38*(6), 477–520.

Ebert, C., C. Ebert, and R. Hörnig (2020). Demonstratives as dimension shifters. In *Proceedings of Sinn und Bedeutung*, Volume 24, pp. 161–178.

Esipova, M. (2019). *Composition and projection in speech and gesture*. Ph. D. thesis, New York University.

Kennedy, C. (2007a). Modes of comparison. In *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, Volume 43, pp. 141–165. Chicago Linguistic Society.

Kennedy, C. (2007b). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy 30*(1), 1–45.

Kennedy, C. and L. McNally (2005). Scale structure, degree modification, and the semantics of gradable predicates. *Language*, 345–381.

Kentner, A. M. (2020). *Examining the Syntax and Semantics of ASL MORE-and BEAT-constructions*. Ph. D. thesis, Purdue University Graduate School.

Klein, E. (1980). A semantics for positive and comparative adjectives. *Linguistics and Philosophy 4*(1), 1–45.

Konrad, R., T. Hanke, G. Langer, D. Blanck, J. Bleicken, I. Hofmann, O. Jeziorski, L. König, S. König, R. Nishio, A. Regen, U. Salden, S. Wagner, S. Worseck, O. Böse, E. Jahn, and M. Schulder (2020). MEINE DGS – annotiert. Öffentliches Korpus der Deutschen Gebärdensprache, 3. Release / MY DGS – annotated. Public Corpus of German Sign Language, 3rd release.

Koulidobrova, E., G. M. Vera, K. Kurz, and C. Kurz (2023). Revisiting gradability in american sign language (asl). *Glossa: a journal of general linguistics 8*(1).

Maier, E. (2018). Quotation, demonstration, and attraction in sign language role shift. *Theoretical Linguistics 44*(3-4), 265–276.

Schlenker, P. (2021). Iconic presuppositions. *Natural Language & Linguistic Theory 39*, 215–289.

Steinbach, M. (2023). Angry lions and scared neighbors: Complex demonstrations in sign language role shift at the sign-gesture interface. *Linguistics 61*(2), 391–416.

Wellwood, A. (2015). On the semantics of comparison across categories. *Linguistics and Philosophy 38*, 67–101.

Wilbur, R. B., N. Abner, S. Wood, and H. Koulidobrova (2018). When BEAT is 'exceed': verbal comparison in American Sign Language. *FEAST. Formal and Experimental Advances in Sign language Theory 1*, 59–69.

Wilbur, R. B., E. Malaia, and R. A. Shay (2012). Degree modification and intensification in American sign language adjectives. In *Logic, Language and Meaning: 18th Amsterdam Colloquium, Amsterdam, The Netherlands, December 19-21, 2011, Revised Selected Papers*, pp. 92–101. Springer.

Zucchi, S. (2017). Event categorization in sign languages. In *Handbook of Categorization in Cognitive Science*, pp. 377–396. Elsevier.