Informational content vs. discourse orientation : Experimental and computational perspectives

Grégoire Winterstein, Ghyslain Cantin-Savoie,
Samuel Laperle, Josiane Van Dorpe & Nora Villeneuve *

Abstract. The aim of this study is to investigate how human speakers and computational language models process (i) the informational content and (ii) the discourse orientation of natural language sentences. These two dimensions of meaning have received little attention outside theoretical literature, especially in the computational linguistics domain. To help fill this void, we present the results of four experiments that exploit the specific semantics of two French adverbs, namely *presque* (≃ 'almost') and *à peine* (≃ 'barely'), which put these two dimensions of meaning at odds. Each experiment focuses on one kind of population (humans or language models), and one kind of meaning (informational content or discourse orientation). Our results show that humans are indeed sensitive to informational content and discourse direction, as assumed in the theoretical literature. Language models exhibit a less transparent behavior. Their performances in dealing with the semantics of *presque* appear in line with predictions based on the way these models are trained, but this does not extend to *à peine*.

Keywords. semantics ; pragmatics ; language models ; psycholinguistics ; entailment ; discourse semantics ; natural langue inference

1. Introduction. A fair number of approaches to natural language semantics consider meaning in terms of *informational content*, understood as the part of an utterance's meaning which conveys information about the world. Informational content is typically understood in truth-conditional terms, and the meaning of a declarative sentence is often defined in such logical terms.

However, natural meaning encompasses more than informational content, at least at a first glance. To see it, consider the (made-up) dialogue (1), taking place in a pub in which patrons need to get their beer at the bar. *A* finished their beer, gets up to get another one, and asks *B*:

(1) *A:* Can I bring you another beer?
   a. *B:* Yes, I'm done with mine.
   b. *B':* Yes, I'm almost done with mine.
   c. *B":* ? Yes, I'm barely done with mine.

Several things are to be noticed in that example. First, (1-a) and (1-b) can be used to answer A's question with the same effect, i.e., to indicate a desire for another drink. This is so, in spite of the fact that those two sentences are at odds in terms of informational content. This is because *being*

*The authors wish to thank the members of the SLIC research group, as well as members of the audience of ELM2 for their input and comments. This work was partially funded by the *Centre de recherche sur le langage, l'esprit et le cerveau (CRLEC-UQAM)*. Authors: Université du Québec À Montréal, Grégoire Winterstein (winterstein.gregoire@uqam.ca) & Ghyslain Cantin-Savoie, (cantin-savoie.ghyslain@courrier.uqam.ca) & Samuel Laperle, (laperle.samuel@courrier.uqam.ca) & Josiane Van Dorpe, (van$_d$orpe.josiane@courrier.uqam.ca) & Nora Villeneuve, (villeneuve.nora@courrier.uqam.ca).

almost done with one's beer entails that one is not done with it (Amaral 2007, Jayez & Tovena 2008). This observation could be accounted for by considering that even though being *almost done* entails not being done, it still conveys that one is proximally close to being done, which would be enough to warrant getting another round.

However, things get more complicated when considering (1-c). Here, the mention of being *barely done* does not support a positive answer to A's question in (1), but rather will support a refusal. Note that being *barely done* seemingly entails being done, admittedly not by a long stretch, but still in a way that is objectively more advanced than when one is *almost done*. Thus, the explanation we sketched above, relying on a notion of proximality, cannot help us here: it seems that considerations of purely informational content cannot fully account for the discourse uses of the expressions at hand.

Such observations form the bedrock of Ducrot and Anscombre's theory of argumentation within language (Anscombre & Ducrot 1983), which serves as inspiration for the present work. Their main claim is that the meaning of an utterance has an *argumentative* component which determines the discourse orientation of that utterance, i.e. how a speaker can use that sentence in a discourse in order to support or refute other propositions. Crucially, the discourse orientation of an utterance can be at odds with its informational content, and the examples in (1) illustrate such cases. This dissociation between the two kinds of content in those examples is directly imputable to the semantics of *almost* and *barely*. These two adverbs belong to the class of "argumentative operators", i.e., natural language expressions whose semantics affect the argumentative orientation of their host utterance.

The purpose of this research is to conduct an experimental investigation of this dichotomy in both the behaviour of human subjects and that of large language models, as used in contemporary applications of natural language processing. Specifically, we want to test the sensitivity of both "populations" (humans and language models) to each type of meaning (informational content vs. discourse orientation) by testing their behavior on sentences that involve argumentative operators like *almost* and *barely* which put those two dimensions at odds. This was done through four distinct experiments involving the French adverbs *presque* and *à peine*, whose semantics is close to English *almost* and *barely*. Each experiment targeted a specific population and type of meaning. Experiments involving humans participants are reported in section 2, and those with language models in section 3. We conclude and discuss future directions for our work in section 4.

2. Experiments: Human participants. The experiments discussed in this section aimed at providing experimental support for the hypothesis that natural language meaning involves (at least) two distinct dimensions of meaning: (i) an "objective" informational content, which encodes descriptive and referential content, and (ii) a discourse orientation (or argumentative orientation) which determines how that sentence can be integrated in a larger discourse.

In a way, those experiments are mostly a form of sanity check: there is already a large body of theoretical work which argues for such distinctions, and the differences seem intuitive enough. There also exists previous work with similar goals. In particular Amaral (2007: Chap. 3) discusses experimental work which also grounds the difference between the truth-theoretic entailments of proximal adverbs like *almost* and *barely* (which she calls the polar component of the adverbs), and the way those sentences can be used in a discourse (which she refers to as the proximal part of

the meaning). Amaral's results are consistent with the ones we present below, though there are differences in the method used, which we highlight whenever relevant.

In section 2.1, we describe the experiment that targets the sensitivity of participants to the informational content of utterances, and section 2.2 introduces the experiment about discourse orientation. We discuss the results of both experiments in section 2.3.

2.1. HUMAN PARTICIPANTS AND INFORMATIONAL CONTENT (EXPERIMENT 1). The goal of the first experiment was to determine whether participants were sensitive to the purported logical entailments of the French adverbs *presque* (≃ 'almost') and *à peine* (≃ 'barely'). Both adverbs modify gradable, therefore scalar, predicates, and so the entailments at hand are as follows:

- an expression of the form *presque X* should indicate a degree of the relevant scale that is lower than the use of *X* alone
- an expression of the form *barely X* should indicate a degree of the relevant scale that is higher than the use *almost X*, and equal or greater than the minimal degree at which *X* is taken to be true

2.1.1. METHOD. The experiment was administered via an online questionnaire, hosted on the `PCIbexFarm` platform (Zehr & Schwarz 2018). 43 participants were presented with a sentence in bold face and asked to place the eventuality described by the sentence along a scale with the use of a slider. The scale was presented below the sentence, and both extrema of the scale were spelt out. The scales were specific to each item, and designed so that the middle of the scale corresponds to the minimal degree such that the predicate used in the target sentence would be true. Figure 1 illustrates an item.

Alex a fini sa bière.

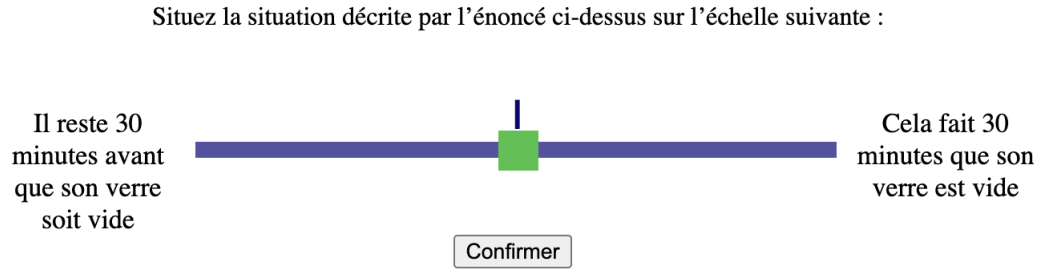

Figure 1: Item example from slider experiment (translations: *Top sentence (in bold)*: Alex almost finished their beer. ; *Instructions*: Locate the situation described by the sentence above on the scale below ; *Left end of the slider* : Alex is 30 minutes away from finishing their glass. ; *Right end of the slider* : Alex finished their glass 30 minutes ago.

On figure 1 the item to be judged is *Alex finished their beer*. For that sentence, the scale is a temporal one, where the extrema are equally distant from the exact moment Alex empties their glass: any place on the right hand part of the scale thus corresponds to an instant when Alex has finished their beer.

The experiment used 18 target items and 18 filler items. We considered three conditions across

target items:

- ∅: the target sentence with an unmodified predicate (as in figure 1)
- `presque`: the target sentence with the predicate modified by *presque*
- `apeine`: the target sentence with the predicate modified by *à peine*

The presentation of items was pseudo-randomized using a latin-square design, so that every participant saw each condition 6 times, with differing orders and condition instantiations across participants.

Participants were recruited via snowball sampling, and each received a link that led them to one of two online questionnaires: either the one for this experiment, or the one for Experiment 2 (described in section 2.3). Participants were told they could only answer the questionnaire once, so that the sets of participants to the two experiments are disjoint.

2.1.2. RESULTS. The average Z-scores (per participant) for each condition in Experiment 1 are presented on figure 2.

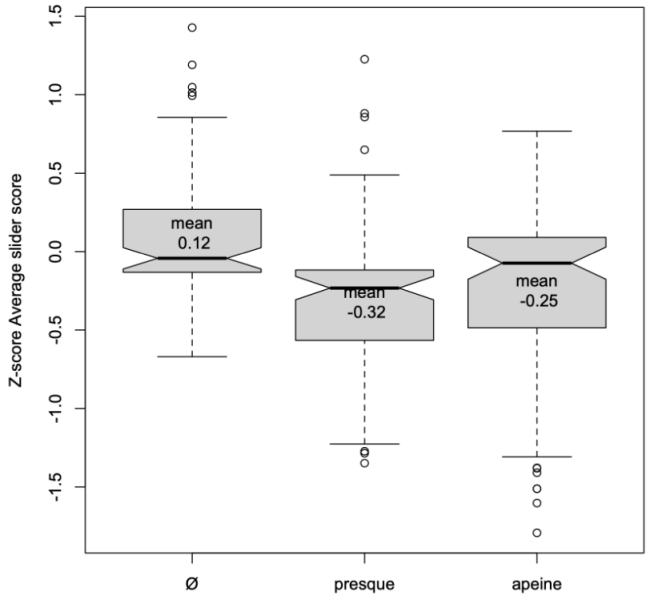

Figure 2: Average Z-scores by condition in Experiment 1

To measure the significance of the condition under study, we fitted linear mixed effect models with random intercepts for items and participants, and assessed the significance of our main factor via model comparison using likelihood ratio tests. We found a significant effect ($\chi^2 = 34.741, p < 0.001$), with the `presque` condition being scored significantly below the other two. The difference between `apeine` and ∅ was not significant.

2.2. HUMAN PARTICIPANTS AND DISCURSIVE ORIENTATION (EXPERIMENT 2). To get an account of the human ability to perceive the argumentative or discourse orientation of utterances, we ran an experiment in which we asked subjects to evaluate the naturality of a sentence in a given context. Regarding our previously discussed hypothesis, our prediction is that the context in which

the sentences with *presque* are judged to be natural are the same contexts in which bare sentences are judged natural, and that in such contexts the use of *à peine* will be judged odd. Conversely, if we change the context in a way which allows *à peine* to be natural, we expect *presque* to sound degraded.

2.2.1. METHOD. As for Experiment 1, Experiment 2 was administered via an online questionnaire, hosted on the `PCIbexFarm` platform (Zehr & Schwarz 2018). 30 participants were presented with a short context, followed by a line break and the target sentence. Participants were asked to rate the naturalness of the sentence in the given context, using a 7 point Likert scale. In (2) we present a (translated) example of target item.

(2) **Context** : Alex and Jackie are enjoying a beer on a terrace of a bar. The waiter comes near them and asks : "Can I bring you another beer?". Alex answers :
Target : <Yes/No>, I'm <almost/barely/$\emptyset$ > done with mine.

As can be seen in (2), we manipulated two different factors. The first was the conclusion targeted by the speaker of the target sentence. In (2), the two possibilities were *Yes* (`Pos` conclusion) and *No* (`Neg` conclusion). The other factor was similar to the one in Experiment 1, i.e. the modification of the predicate in the target sentence (with three levels $\emptyset$, `presque` and `apeine`). This created 6 different conditions in total. The experiment included 18 target items and 36 distractors. The presentation of items was pseudo-randomized using a latin-square design, so that every participant saw each condition 3 times, with differing orders and condition instantiations across participants. Participants were recruited in the same manner and at the same time as for Experiment 1 (see section 2.1.1 above).

2.2.2. RESULTS. The results of Experiment 2 are summarized in Figure 3. Model comparison between ordinal mixed models with random intercept and slopes for items and participants shows a significant effect of each of the factor under study, i.e. of the modification of the predicate ($\emptyset$/`presque`/`apeine`, $\chi^2 = 34.98, p < 1e^{-8}$) and the type of conclusion target (`Pos`/`Neg`, $\chi^2 = 18.77, p < 1e^{-5}$). The interaction between the two factors is also significant ($\chi^2 = 160.65, p < 1e^{-10}$). Overall, the acceptability of context favoring positive conclusions was higher, but within each condition, preferences were reversed: within `Pos` contexts, $\emptyset$ patterned with `presque`. In `Neg` contexts, `apeine` was judged significantly more natural than the two other conditions, but some discrepancies were observed between $\emptyset$ and `presque`.

2.3. DISCUSSION. The results of Experiments 1 and 2 largely support the predictions of theoretical models on the interpretation of proximal adverbs and are in line with previous experimental work in this domain. Comparing our experiments with those of Amaral (2007), we mostly differ in how we tested the sensitivity of participants to the entailments of the target sentences. While Amaral asked participants for binary entailment judgment, our method relied on graded behaviors in Experiment 1. This allowed us to observe that for certain items, the degree associated to $\emptyset$ was close to that of `apeine` while in other cases the degree of $\emptyset$ appeared higher than `apeine`'s. Both configurations are compatible with a truth-conditional entailment pattern of *à peine* to its prejacent, but not necessarily one in terms of degrees, i.e. though *à peine* might be felt to entail the truth of its prejacent, it does not entail that the prototypical degree associated to the property holds

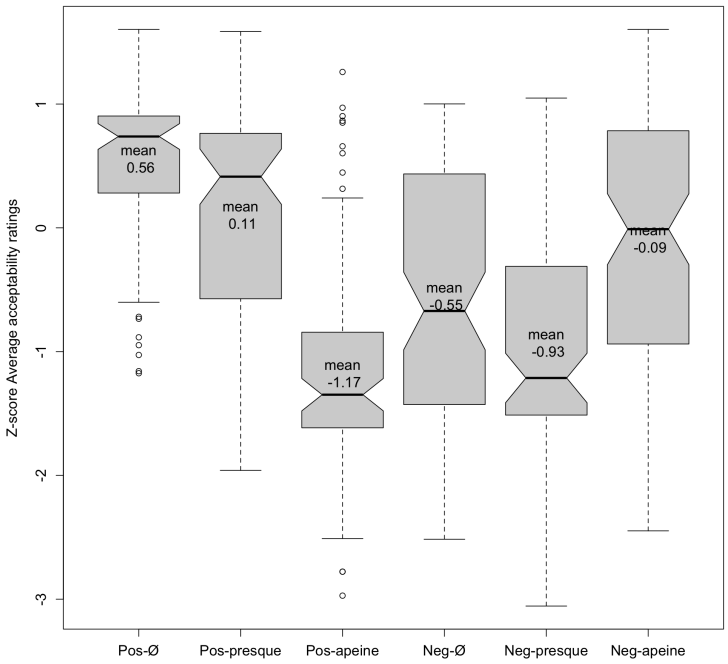

Figure 3: Average Z-scores of naturality judgments by condition in Experiment 2

(which corresponds to the minimizing effect of *à peine*).

Overall, our results are thus consistent with the hypothesis that the linguistic knowledge of participants encompasses both the informational content of the utterances and their discourse orientation, and that they are able to tease those apart.

3. Experiments: Language models. Computational language models are often taken to encode complex grammatical knowledge. In particular, contemporary models such as Transformer based models (Vaswani et al. 2017) (e.g. those of the BERT family, Devlin et al. 2019) are typically pre-trained on large amounts of data on relatively general tasks (such as predicting the nature of a masked token). This pre-training is supposed to capture general linguistic knowledge, which can later be fine-tuned for particular tasks.

The pre-training of the models crucially relies on distributional information: the representations encoded by the models are rooted in the observation of co-occurrence patterns in the training data. In that sense, we expect the models to have captured information related to the discourse orientation of linguistic elements, since discourse orientation precisely relates to matters of co-occurrence at the discourse level. Language models are however often fine-tuned for tasks like the *Natural Language Inference* (NLI) task in which the model is used to predict whether one sentence, called the premise, entails another, called the hypothesis (Poliak 2020). Such a task would thus rely on the manipulation of informational content rather than discourse orientation. Experiments 1 and 2 are consistent with the hypothesis that human speakers are sensitive to, and can distinguish between, discourse orientation and informational content. We designed two additional experiments

to test whether language models are also sensitive to both dimensions, using the same adverbs as in the experiments with human participants.

We begin by describing how we built the dataset used in both experiments (section 3.1). We then introduce Experiment 3 in which we test the predictions of fine-tuned language models on detecting the inferential patterns associated with each adverbs (section 3.2). In section 3.3, we present Experiment 4 which targets the sensitivity of language models to discourse orientation. We discuss the results of these experiments in section 3.4.

3.1. DATASET. Our dataset consisted in naturally occurring sentences containing one of the two adverbs under study.

The dataset was built by pseudo-randomly extracting sentences from a subset of the French version of Wikipedia which contained either *presque* (1990 items) or *à peine* (2770 items). In addition to the target sentence, we also extracted the two preceding sentences to serve as context.

Every target sentence was then replicated in three conditions:

- `Original`: the original extracted sentence
- `Bare`: the sentence with the target adverb removed
- `Switched`: the sentence with the target adverb replaced by the other (i.e. *presque* replaced by *à peine* and vice-versa)

Every sentence was also tagged with the adverb that actually appears in it (irrespective of its original form), with the same possible values as in experiment 1 and 2, i.e. `∅/presque/apeine`.

3.2. LANGUAGE MODELS AND INFORMATIONAL CONTENT (EXPERIMENT 3). To test the sensitivity of language models to informational content, we evaluated a model that was previously fine-tuned to the NLI task on pairs of sentences from our dataset. If the model acquired some knowledge about the informational content of the adverbs under study, we expect it to predict that:

- a sentence containing *presque* should contradict its prejacent
- a sentence containing *à peine* should entail its prejacent

3.2.1. METHOD. For the experiment, we used the `CamemBERT` model for French (Martin et al. 2020) which is pre-trained on french data and fine-tuned on the French part of the MNLI dataset (Williams et al. 2018). We tested the entailment patterns by taking `Original` sentences as premises, and the `Bare` sentences as hypothesis. For each premise/hypothesis pair the model gave us the probability that the hypothesis is true given the premise.

3.2.2. RESULTS. The results of the experiment are summarized in Figure 4. As can be seen on the figure, the entailment patterns of the two adverbs starkly differ. For sentences containing *presque* the general prediction is that they entail their prejacent. On the other hand, sentences with *à peine* appear to be divided in two groups: those that entail their prejacent and those that do not, with few sentences in between.

3.3. LANGUAGE MODELS AND DISCURSIVE ORIENTATION (EXPERIMENT 4). To test the sensitivity of language models to discursive orientation we relied on pre-trained models, before any task-specific fine-tuning. The rationale of our choice is that the pre-training of language models aims at capturing matters related to the distribution of linguistic expressions, and that discursive

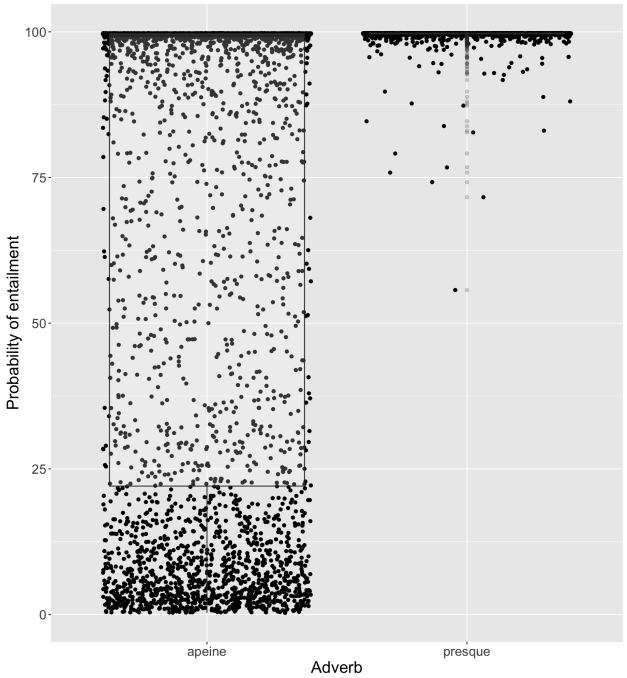

Figure 4: Probability of entailment of the prejacent for sentences containing the target adverbs

orientation boils down to a question of distribution. In Experiment 2, we used the naturality of discourses to assess the sensitivity of human participants to discursive orientation. Here, we will use the perplexity of a model towards the target sentences, a measure we define in the next subsection.. If the model acquired some information about the discursive orientation of the adverbs under study, we expect it to predict that:

- replacing *presque* by *à peine*, or vice-versa should increase the perplexity
- deleting *presque* should not significantly alter the perplexity, or do it in a way not as marked as a deletion of *à peine*

3.3.1. METHOD. To test the sensitivity of language models to discourse orientation, we relied on the the `French-GPT` model (Simoulin & Crabbé 2021). The rationale for our choice was that, unlike BERT models, GPT-like models are directional, and thus are suited to measuring the perplexity of the model about the continuation of a discourse. As mentioned above, we take the perplexity of the model as a proxy for the naturality of a discourse: the higher the perplexity, the less natural the discourse.

We thus calculated the average perplexity of the model on our target sentences, in the three conditions `Original`, `Switched` and `Bare`, using the two previous sentences as the context against which to measure the perplexity of the model. The perplexity measure uses the context before a particular word to assess its probability of occurrence. Specifically, it is defined as *"the exponentiated average negative log-likelihood of a sequence. If we have a tokenized sequence* $X = (x_0, x_1, \ldots, x_t)$*, then the perplexity of XX is defined [as follows], where* $\log p_\theta(x_i|x_{<i})$ *is the*

log-likelihood of the i-th token conditioned on the preceding tokens $x_{<i}$"[1].

$$\text{PPL}(X) = e^{-\frac{1}{t}\sum_{t}^{i}\log p_\theta(x_i|x_{<i})} \tag{1}$$

3.3.2. RESULTS. Figure 5 summarizes the results of the experiment by showing the average Z-score for perplexity in each of the three conditions, for each adverb that was present in the `Original` sentences (e.g. the bar for `switch/apeine` represents cases in which the original *à peine* was replaced by *presque*).

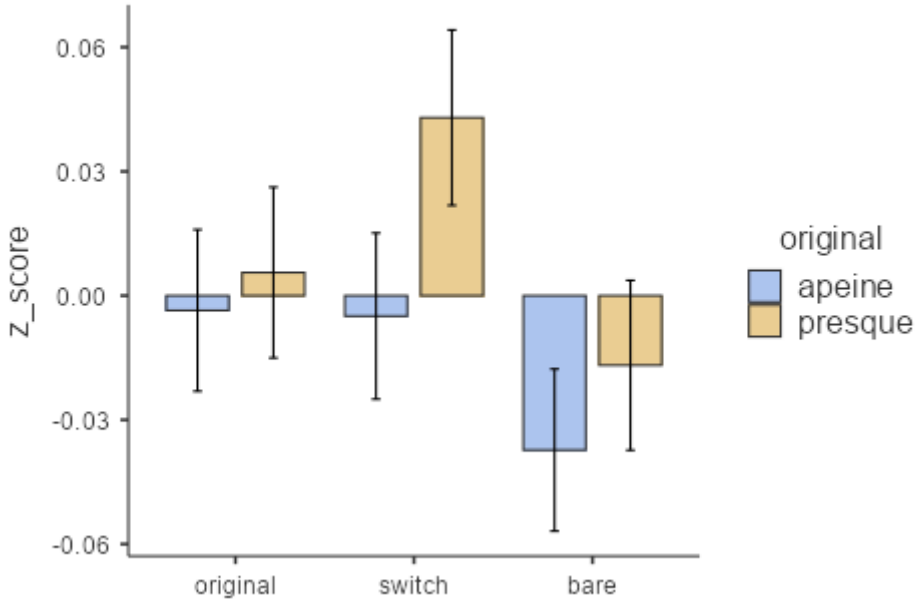

Figure 5: Z-scores for perplexity across conditions in Experiment 4.

The general patterns we can observe is that replacing *presque* by *à peine* will greatly augment the perplexity of the model, and that removing the adverbs will lower the perplexity, especially in the case of *à peine*.

3.4. DISCUSSION. Overall, the results of Experiments 3 and 4 suggest different conclusions.

For *presque*, Experiment 3 shows that a model specifically trained for natural language inference will nevertheless predict the wrong entailment pattern, namely that *presque* entails its prejacent. The results are more subtle for *à peine*, though they do make sense in the light of the results of Experiment 1. These results showed that while human participants systematically rated the `presque` condition below the ∅ and the `à peine` conditions, the rankings of ∅ and `à peine` were not clear-cut. In other words, there were situations in which participants put the prototypical degree of ∅ above that of `à peine`. The entailment patterns of *à peine* in Experiment 3 could thus be interpreted as reflecting those two options. When the degree of ∅ significantly exceeds that of `à peine`, non-entailment is predicted, while in other cases, entailment would be the prediction.

[1] https://huggingface.co/docs/transformers/perplexity

For discourse orientation, the results are not fully consistent with the hypothesis that the models are sensitive to the different orientations of *presque* and *à peine*. While substituting *à peine* for *presque* does increase the perplexity, as expected, the reverse substitution does not seem to alter the perplexity in any significant way. This suggests that it is *à peine* itself which is responsible for a high perplexity. Indeed, if we compare the `Bare` cases with the `Original` ones for *à peine*, we find that removing *à peine* significantly reduces the perplexity. This might possibly be related to the fact that *à peine* is a two words expression, the first of which, *à*, is an extremely frequent preposition in French. Overall, the appearance of *à* in *à peine* is a rarer occasion than its other uses, which in turn will affect the perplexity of the model when processing this element.

4. General Discussion and conclusion. Overall, and as expected, the results of Experiments 1 and 2 largely support previous findings and theoretical claims about the semantics and interpretation of *presque* and *à peine*. French speakers (and most likely speakers of other languages) are sensitive to both the informational content entailed by those adverbs and the way these adverbs can be used in a discourse. Depending on the task at hand, participants exhibited different behaviors, pairing *almost* with its prejacent in terms of discourse orientation, but teasing them apart for matters of entailment, with a mirror behavior for *à peine*. This is thus consistent with the hypothesis that speakers distinguish between these two dimensions of meaning, using them as the basis of distinct processes.

On the other side, the results of our experiments on language models are less straightforward to interpret. Though we can account for some of the wrongful entailment patterns by making the hypothesis that such models learn from distributional data, and are thus more sensitive to discourse orientation, other patterns cannot be fully accounted for by that explanation. One possible explanation rests on the fact that *à peine* is a two-words expression, which might act as a confound on the measure of perplexity, which we use as an approximation of discourse naturality for language models.

To address that point, future work will go in two complementary directions. First, we intend to run comparable experiments in languages other than French, starting with English. In many of its uses *à peine* is comparable to English *barely*, which consists of only one word, and would thus not entail the same issues as *à peine*. Second, we also intend to widen the range of argumentative operators under study. The choice of *presque* and *à peine* was motivated by their apparent dual nature, especially in the way they put their informational content and discourse orientation at odds. These two adverbs are however not the only ones to do so: exclusive adverbs like *only* and *just* also reverse the orientation of their host sentence while still conveying its content (Winterstein 2012). We will thus extend our computational experiments to more operators, both similar in profile to *presque* and *à peine*, but also that exhibit other configurational patterns between their entailed content and discourse orientation.

References

Amaral, Patricia. 2007. *The meaning of approximative adverbs: Evidence from European Portuguese*: The Ohio State University dissertation.

Anscombre, Jean-Claude & Oswald Ducrot. 1983. *L'argumentation dans la langue*. Liège, Bruxelles: Pierre Mardaga.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In Association for Computational Linguistics (ed.), *Proceedings of NAACL-HLT 2019*, 4171–4186.

Jayez, Jacques & Lucia Tovena. 2008. Presque and almost: how argumentation derives from comparative meaning. In Olivier Bonami & Patricia Cabredo Hofherr (eds.), *Empirical Issues in Syntax and Semantics*, vol. 7, 1–23. CNRS.

Martin, Louis, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah & Benoît Sagot. 2020. Camembert: a tasty french language model. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics(acl*, .

Poliak, Adam. 2020. A survey on recognizing textual entailment as an NLP evaluation. In *Proceedings of the first workshop on evaluation and comparison of NLP systems*, 92–109. Online: Association for Computational Linguistics. 10.18653/v1/2020.eval4nlp-1.10. https://aclanthology.org/2020.eval4nlp-1.10.

Simoulin, Antoine & Benoit Crabbé. 2021. Un modèle Transformer Génératif Pré-entrainé pour le français. In Pascal Denis, Natalia Grabar, Amel Fraisse, Rémi Cardon, Bernard Jacquemin, Eric Kergosien & Antonio Balvet (eds.), *Traitement Automatique des Langues Naturelles*, 246–255. ATALA. https://hal.archives-ouvertes.fr/hal-03265900.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser & Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett (eds.), *Advances in Neural Information Processing Systems (anips)*, vol. 30, Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Williams, Adina, Nikita Nangia & Samuel 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)*, 1112–1122. New Orleans, Louisiana: Association for Computational Linguistics. 10.18653/v1/N18-1101. https://aclanthology.org/N18-1101.

Winterstein, Grégoire. 2012. "Only" without its scales. *Sprache und Datenverarbeitung* 35-36. 29–47.

Zehr, Jeremy & Florian Schwarz. 2018. PennController for internet based experiments (IBEX). https://doi.org/10.17605/OSF. IO/MD832.