Towards a Formal Theory of Explanatory Biases in Discourse

Torgrim Solstad[1] and Oliver Bott[2]

[1] Department of Language and Literature, Norwegian University of Science and Technology
torgrim.solstad@ntnu.no
[2] SFB 833, University of Tübingen
oliver.bott@uni-tuebingen.de

Over the past four decades, psycholinguistic studies of discourse structure have revealed interpretation biases for a class of verbs called *implicit causality* (henceforth, IC) verbs. In brief, IC verbs are transitive verbs with two animate arguments which show a preference for anaphorically referring to one of the two arguments when followed by a *because* sentence, cf. (1):

(1) a. ***fascinate*, NP1 bias:** *John* fascinated Mary because *he* danced beautifully.
b. ***congratulate*, NP2 bias:** John congratulated *Mary* because *she* won the race.

Such preferences are commonly elicited in production experiments where participants are prompted to continue sequences such as *"John congratulated Mary because ..."*. The proportion between continuations mentioning the subject or object argument first is referred to as *IC bias*. Referring to the linear order of arguments, anaphoric reference to the subject argument contributes to NP1 bias, whereas reference to the object argument contributes to NP2 bias.

Although highly consistent across production and comprehension, experimental paradigms and languages, IC bias has attracted little interest in semantic and pragmatic theory. Previous approaches have focused on the establishment of correlations between reference resolution and thematic properties of arguments (e.g. [4, 6]). For instance, it has been shown that the stimulus arguments of psychological verbs are strong bias attractors, cf. (1a). However, little is known as to *why* it is that certain thematic roles attract the bias (and others don't).

IC verbs are special in other respects, too. In a recent paper, [7] have shown that IC verbs are prone to trigger explanations in subsequent discourse. Participants were prompted to continue sequences such as *"John congratulated Mary."*, i.e. after a full stop. In such cases, IC verbs triggered 60% explanations, i.e. sentence continuations which where causally related to the prompt, whereas "non-IC verbs" triggered only 25% explanations. However, [7] offered no explanation for this property.

This, however, is a rather interesting property, in particular because it promises to broaden the scope of discourse-theoretic studies. Consider the mini-discourse in (2):

(2) [John congratulated Mary.]$_{s_n}$ [She won the race.]$_{s_{n+1}}$

In (2), s_{n+1} is most naturally interpreted as an explanation of s_n, cf. (1b). For such unmarked discourse relations, i.e. without a connective, discourse-theoretic studies have been concerned with the *post hoc* processing of discourse, asking how s_{n+1} is integrated into s_n. From the point of view of IC, however, one would ask already after processing s_n which discourse relation could be *expected* to be established with the next segment. In our study of IC, we investigate lexical triggers of such expectations. We will point at two well-established phenomena from semantic theory involved in such a forward-looking perspective (cf. [8] for experimental evidence).

In this paper, we propose a semantic theory of IC which incorporates the discourse coherence and reference resolution properties of IC verbs, explaining why IC verbs pattern the way

they do. Crucially, our approach links the coreference patterns to the observed preference for explanations [7]. We also present comprehensive experimental evidence in favour of the theory with important implications for future experimental research.

A Semantic Theory of Impiclit Causality

IC verbs, we contend, trigger expectations for *specific explanation types*. They do so because they are underspecified with respect to certain properties of the situation described which are (causally) contingent on one of the two participants. Put differently, IC verbs carry an empty "slot" for specific explanatory content. It is this missing information which triggers explanations in full stop continuations and which also triggers primary reference to one of the two participants. IC thus reflects a general processing preference for not leaving missing content unspecified, i.e. a tendency to avoid accommodation [2, 10]. On our analysis, IC bias as a measure of coreference preferences is an epiphenomenon of specific explanatory preferences derived from verb semantics and the particular realization of its arguments.

In order to capture the relation between explanations and coreference, we need to distinguish several types of explanations. Based on [9], we distinguish three main types of explanations, (i) simple causes, (ii) external reasons, and (iii) internal reasons, cf. the examples in (3):

(3) a. **simple cause**: *John* disturbed Mary because *he* was making lots of noise.
b. **external reason**: John disturbed *Mary* because *she* had damaged his bike.
c. **internal reason**: *John* disturbed Mary because *he* was very angry at her.

Simple causes are causes of events or (mental) states. They never involve volition or intention. In (3a), Mary feeling disturbed is understood to be a by-product, as it were, of John's noise-emitting activity. *External and internal reasons* are causes of attitudinal states. Thus, the *because* clauses in (3b)-(3c) specify causes for John's intention to disturb Mary. External reasons (3b) are states of affairs external to the attitude-bearer's mind whereas internal reasons (3c) are attitudes or mental states internal to the attitude-bearer's mind. The interdependency between explanation type and reference resolution may be seen in (3b)-(3c). The external reason is associated with, and thus makes primary reference to the object (NP2) argument by mentioning it first, whereas the internal reason is associated with the subject (NP1) argument.

We will also assume *backgrounds* as a fourth type of explanation. Backgrounds provide information which make possible, or 'facilitate' the situation described by the verb, cf. (4):

(4) a. Felix frightened Vanessa because he suddendly screamed.
b. Felix frightened Vanessa because she didn't hear him coming.

Whereas the *because* clause in (4a) specifies the simple, direct cause of Vanessa's state of being frightened, the *because* clause in (4b) specifies the background (or: preconditions) for Vanessa being frightened, saying nothing about the actual cause of Vanessa's fear.

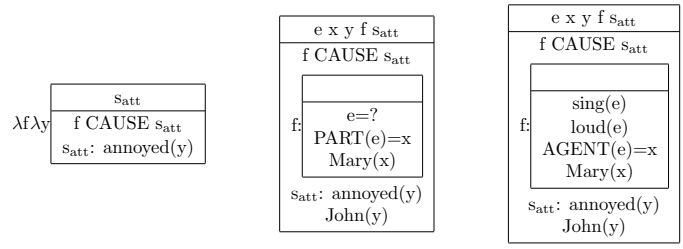
Finally, it is of importance to our approach that *because* clauses introduce entities propositional in nature, cf. [9]. Thus, if *because* clauses are to suitably specify underspecified causal entities, these must likeweise be of a propositional semantic type.

Two kinds of underspecified content trigger explanatory expectations. One involves arguments which are underlyingly propositional in nature. Consider the stimulus argument of the stimulus-experiencer verb *annoy*. In the sentence *Mary annoyed John*, *Mary* may be seen as a mere placeholder of a semantic entity more complex in nature. It is actually a specific property or action of Mary's which is the cause of John being annoyed. Support for this analysis derives from the fact that stimuli in general may be realized as either noun phrases or *that* clauses:

Mary *annoyed John/It annoyed John* ***that Mary ...*** Stimuli are simple causes, contributing to NP1 bias for stimulus-experiencer verbs, and to NP2 bias with experiencer-stimulus verbs.

Although we cannot go into great detail, we will show by way of example how one can conceive of our "missing content" approach to IC. Consider the DRSs in (5), which illustrate an incremental construction of *Mary annoyed John because she sang loudly*:

(5)



The left-most DRS (A.) shows a possible lexical entry for *annoy*. Crucially, the subject argument of *annoy* is associated with a discourse referent f which is propositional in nature. When this argument is realized by a DP introducing an entity, this leads to a reinterpretation of the entity argument, much akin to complement coercion (cf. also the mechanism of type-shifting assumed for concealed questions in [1]). A first step in this process is shown in the middle DRS (B.): It is not specified which event involving Mary (it could also be a property of hers) it is that causes annoyance in John (which is what the brief-hand notation *e=?* represents). All we know is that Mary is a participant in this event. Finally, the right-most DRS shows the result of modifying the sequence *Mary annoyed John* by means of the *because* clause *because she sang loudly*.

The other trigger of explanation expectations which we have identified involves presuppositions suitable to give an external reason for the execution of the action denoted by an agent-patient verb (these verbs have previously been assigned to the *ad hoc* class of 'agent-evocator' predicates). Take *congratulate*. For it to be used adequately, there must be some occasion on which the agent may congratulate the patient, e.g. an event occurring prior to the act of congratulating. It is exactly this 'occasioning circumstance' which is specified by means of the *because* sentence: *John congratulated Mary because she won the race*. Importantly, this is a presupposition which is – for lack of a better term – *cataphorically verifiable*, i.e. after the occurence of the trigger. Such presuppositions are external reasons, contributing to NP2 bias.

For verbs which lack a relevant underspecified slot, such as agent-patient verbs without an "explanatory" presupposition or causative verbs other than stimulus-experiencer verbs, there is no relevant causal content missing. This basically leaves external and internal reasons as alternative explanations. For some such verbs, a strong bias may still be observed. We cannot yet predict the bias of these verbs on semantic or pragmatic grounds. However, if we are right about the bias resulting from the explanatory strategy sketched above, the bias is predicted to follow from the ratio of external to internal reasons. This is shown in Experiment 2.

Experimental evidence

Testing our semantic theory of IC, we conducted two discourse continuation experiments in German and Norwegian. In Experiment 1a/b, we tested the effects of specifying the bias-triggering missing causal content of psych verbs (Experiment 1a) and presupposition verbs (Experiment 1b) by means of a modifier in the matrix clause. By introducing the modifier, we preempt the strategy of avoiding accommodation which we argue to be behind the observed discourse continuation strategy. If we are right about this, we expect the following effects: (i) a drop in the proportion of explanations of the expected kind, and consequently, (ii) a shift

in IC bias. In Experiment 2, we show that our assumptions are cross-linguistically valid in a large-scale study of German and Norwegian.

Experiment 1a: Modification by means of *durch* 'by' Phrases. In this experiment, we specified the expectation-triggering, missing content of stimulus experiencer verbs by means of an adverbial modifier, cf. (6a) (adverbial modifier in italics):

(6) a. Peter faszinierte Linda (*durch seine Reiseberichte*) a), weil b) .
Peter fascinated Linda (*by/with his travel-descriptions*) a) because b) full stop
b. Maria tötete Johann (*durch einen Schuss*) a), weil b) .
Mary killed John (*with a shot*) a) because b) full stop

To control for the general effects of *durch*-modification, we also included causative a-p verbs like *kill* which can also be modified by *durch* phrases, cf. (6b). This is crucial, because the causing entity of causative verbs other than stimulus-experiencer verbs is assumed to be an event. Consequently, *because* clauses cannot specify the causing event for these predicates since they introduce causes propositional in nature:

(7) #Mary killed John because she stabbed him in the back.

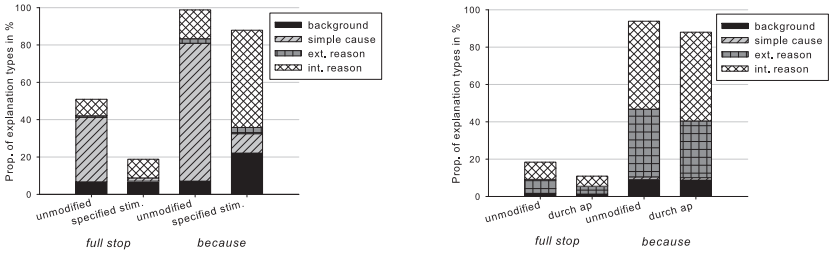
The *because* sentence in (7) cannot be taken to introduce a simple cause as in (6a). Due to this ontological restriction of *because*, the *durch* phrase modification should neither affect the distribution of explanation types nor the bias in the case of causatives such as *kill*.

Methods: 48 native German participants (mean age 23.7 years; range: 19 – 37 years; 33 female) took part in the experiment which employed a $2 \times 2 \times 2 \times 2$ design. We constructed 20 items with s-e verbs and 20 items with causative a-p verbs such as (6a) and (6b). Within each *verb type*, we manipulated the factors *modification* (*durch PP* vs. *no modification*), *connective* (*because* vs. *full stop*) and *gender* ($NP1_{fem.}NP2_{masc.}$ vs. $NP1_{masc.}NP2_{fem.}$). The latter manipulation was included to control for potential effects of gender (see e.g. [5]). The items were presented together with 90 fillers in four lists which were constructed according to a latin square design (only taking into account *verb type*, *modification* and *connective*). The experiment consisted of two blocks. In the first block, participants provided continuations after a full stop (for 20 out of 40 items and 45 fillers). This way, we made sure there was no overt bias towards a particular coherence relation. In a second block, they were prompted to continue the other items and 45 fillers with *weil* 'because'. Participants who felt that no sensible continuation was possible, had the option to press a "no sensible continuation" button.

Corpus annotation: We removed all nonsensical and ungrammatical continuations from the corpus (3.5% of the data). Continuations after a full stop were annotated with respect to discourse relations using a *because* insertion test to identify explanations (cf. [7]). Explanations after a full stop and the continuations in the *because* conditions were annotated with respect to the causal typology described above; being categorized as *simple causes* (*SC*), *external reasons* (*ER*), *internal reasons* (*IR*) or *backgrounds*. Finally, we annotated the IC bias in all continuations which were explanations. Details on the annotation scheme can be found in [3].

Results and discussion: In the following, we will first report the results for s-e verbs and then present the results obtained for a-p verbs. For statistical analysis we computed log-linear models analyses including factors for participants and items. In the following we report the partial associations. Note that there is no distinction between independent and dependent variables within these models. $LRCS_1$ (LRCS=log-likelihood ratio χ^2) refers to the analysis by subjects, $LRCS_2$ to the analysis by items.

s-e verbs: Fig. 1a shows the proportions of explanations (for each type) that were produced



(a) *s-e* verbs (b) *a-p* verbs

Figure 1: Types of explanations in Experiment 1a.

after full stop and *because* prompts. Trials in the *because* conditions judged not to be continuable in a sensible way were counted as non-explanations. After a full stop, participants produced explanations in 51.3% of the unmodified cases, but only in 19.2% of the modified prompts. A drop in the proportion of explanations was also observed in the *because* conditions where participants provided continuations in 99.2% of the unmodified cases. In the *durch* condition, however, they were unable to produce a continuation in 6.2% of the cases.

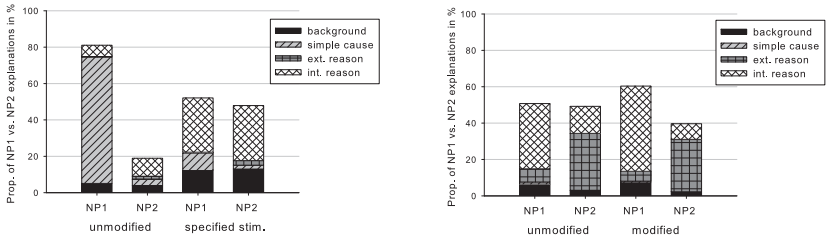
The statistical analysis revealed significant partial associations between *modification* and *explanation* ($LRCS_1(1) = 72.9, p < .01; LRCS_2(1) = 69.1, p < .01$) and *connective* and *explanation* ($LRCS_1(1) = 525.8, p < .01; LRCS_2(1) = 494.4, p < .01$) but no significant interaction between *modification*, *connective* and *explanation* ($LRCS_1(1) = .85; LRCS_2(1) = 1.5$).

The drop in explanations was accompanied by a shift in the type of explanation. Explanations in the unmodified conditions were predominantly simple causes (68.6% *SC* relative to other explanation types after a full stop; 75.0% after *because*). In the modified conditions, however, *internal reasons* were the default explanation type (modified full stop condition: 53.3% IR; modified *because* condition: 59.2%). Statistical analysis revealed that the shift in explanation type was significant (interaction between *modification* and *SC vs. ER/IR/other*: $\text{LRCS}_1(1) = 284.3, p < .01; \text{LRCS}_2(1) = 266.6, p < .01$).

As predicted, the shift in explanation type led to a shift in IC bias. Fig. 2a presents the proportion of continuations that were explanations coreferent with NP1 or NP2 and the explanation profiles of both NP1 and NP2 continuations. Simple causes refer to the stimulus argument leading to a NP1 bias (81.1% NP1) in the unmodified cases. In the modified cases the bias disappears. Continuations are balanced between NP1 and NP2 coreference (52.1% NP1). Statistical analysis showed that the observed shift in bias was reliable (interaction between *modification* and *bias*: $\text{LRCS}_1(1) = 85.2, p < .01; \text{LRCS}_2(1) = 72.2, p < .01$).

a-p verbs: Fig. 1b shows the proportions of explanations that were produced after full stop and *because* prompts. After a full stop, participants produced explanations in 18.8% of the unmodified cases with a slight drop to 11.0% explanations after *durch* phrases. Statistical analysis also revealed significant partial associations between *modification* and *explanation* ($\text{LRCS}_1(1) = 8.1, p < .01; \text{LRCS}_2(1) = 7.1, p < .01$). However, a loglinear model comparing the relative drop in explanations for the verb types revealed a significant interaction between *verb type*, *modification* and *explanation* ($\text{LRCS}_1(1) = 6.9$, $p < .01$) which was due to the fact that the drop in the proportion of explanation was stronger for the *s-e* than the *a-p* verbs.

As predicted, the distribution of explanation types was the same for modified and unmodified *a-p* verbs as indicated by far from significant partial associations (interaction *modification*×*explanation type* including all four types in the analysis: $\text{LRCS}_1(3) = 1.6; \text{LRCS}_2(3) = .8$).



(a) *s-e* verbs (b) *a-p* verbs

Figure 2: IC bias of explanations in Exp. 1a (aggregated over full stop and *because* conditions).

Fig. 2b presents the proportion of continuations that were explanations coreferent with NP1 or NP2 and the explanation profiles of both kinds. While the unmodified conditions were almost unbiased (50.8% NP1), *durch* modification led to a weak NP1 bias (60.4% NP1). This change was reflected by a significant interaction between *bias* and *modification* in loglinear analyses ($\text{LRCS}_1(1) = 4.8, p < .05; \text{LRCS}_2(1) = 3.9, p < .05$). Crucially, this shift is in the opposite direction of the *s-e* verbs. An explanation may be that in the modified conditions the salience of NP1 might have been higher than in the unmodified conditions since the simple cause provided by the *durch* phrase made implicit reference to the subject. For instance, in sentence (6b) *durch einen Schuss* may strengthen the activation of the subject because it was the agent of the shooting event. The opposite behavior of *s-e* as compared to *a-p* verbs with respect to IC bias was statistically confirmed by a significant interaction between *verb type*, *bias* and *modification* in a loglinear analysis copmparing both verb types ($\text{LRCS}_1(1) = 73.1, p < .01$).

Experiment 1b: Implicit Arguments of PSP verbs. Verbs that carry a presupposition (PSP) such as *congratulate* constitute the second class verbs discussed above. The verification of the presupposition can take various forms. The presupposed external reason can be stated in a *because* clause, but it can also be specified by a prepositional object, as in (8). The experiment shows that providing an external reason in the matrix sentence shifts the bias from NP2 to NP1.

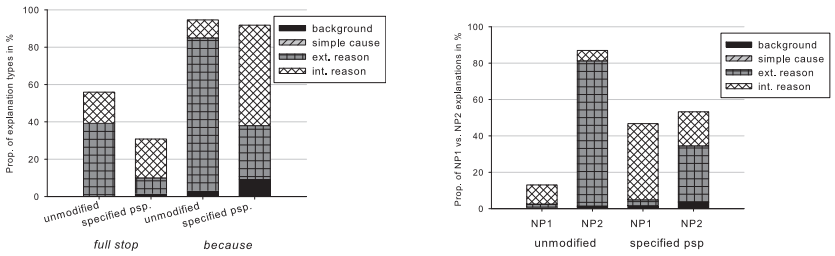
(8) Peter dankte Maria (*für die finanzielle Unterstützung*) a), weil b) .
Peter thanked Mary (*for the financial support*) a) because b) full stop.

Methods: 20 items with PSP verbs were constructed in eight conditions manipulating the factors *modification* (*implicit argument* vs. *no modification*), *connective* (*because* vs. *full stop*) and *gender* ($NP1_{fem.}NP2_{masc.}$ vs. $NP1_{masc.}NP2_{fem.}$). The methods were as in Exp. 1a.

Results and discussion: Fig. 3a presents the proportions of explanation types. Again, specifying the missing causal content led to a significant drop in the proportion of explanations from 55.8% to 30.8% after a full stop and from 94.6% to 91.7% after a *because* prompt (interaction of *modification* and *explanation*: $LRCS_1(1) = 34.3, p < .01; LRCS_2(1) = 31.9, p < .01$).

The types of explanations were as expected. In the unmodified conditions, external reasons were the default (unmodified full stop condition: 70.1% of all explanations were ER; unmodified *because* condition: 86.3% ER). After matrix sentences with a specified presupposition, the proportion of ERs dropped and IRs became the predominant type of explanation (unmodified full stop condition: 67.6% IR; unmodified *because* condition: 58.6% IR). Statistical analyses revealed that this shift in explanation type was significant (interaction between *modification* and *ER vs. SC/IR/other*: $LRCS_1(1) = 193.6, p < .01; LRCS_2(1) = 202.1, p < .01$).

Fig. 3b shows that this led to a reversal in IC bias. Explanations in the unmodified conditions



(a) Proportions of explanations (b) IC bias of explanations

Figure 3: Explanation types and IC bias in Experiment 1b.

displayed a clear NP2 bias (full stop: 84.4% NP2; because: 88.4% NP2). By contrast, conditions that specified the PSP in the prompt, revealed no clear bias (full stop: 44.4% NP2; because: 50.8% NP2) – a highly significant shift (interaction *modification*×*bias*: $LRCS_1(1) = 93.5, p < .01; LRCS_2(1) = 93.4, p < .01$). The analysis of the NP1 and NP2 explanation profiles shows that external reasons are associated with NP2 reference, whereas internal reasons are associated with NP1 reference. Therefore, IC bias can be directly read off the explanation profile lending support to our claim that it is an epiphenomenon of the explanatory strategy.

Experiments 1a/b provide evidence for our claim that specifying missing causal content minimizes the need for explanation, which in turn gives rise to a shift to another discourse relation. We have seen that the discourse processor tends to elaborate on this content probably in order to avoid accommodation costs [2]. Furthermore, the triggered explanations were of the type expected on our analysis. Considering argument associations provided us with a direct way to account for the observed bias distributions. Thus, IC bias seems to be an epiphenomenon of developing a discourse expectation on the basis of presupposed causal content.

Experiment 2 – Large-scale corpus. In this continuation experiment, we elicited 10.100 written productions for 101 near-synonymous German and Norwegian verbs (for details see [3]). 52 German and 48 Norwegian participants were prompted to continue sentences after either a full stop or *because*. Besides 16 stimulus-experiencer and 10 presupposition verbs the study included 17 experiencer-stimulus verbs and 43 agent-patient verbs without a presupposition as well as 14 ambiguous agent-patient/stimulus-experiencer verbs such as *disturb* and *hurt*.

A verb-by-verb comparison revealed that the biases were highly correlated between the two languages ($r = .92$), showing that the phenomenon is cross-linguistically stable as expected under a semantic account. Moreover, our analysis of stimulus arguments generalized from stimulus-experiencer to experiencer-stimulus verbs. Again, we observed explanations to be the default coherence relation. While agent-patient verbs without a presupposition had a mean proportion of only 39.1% explanations, experiencer-stimulus had 63.2% explanations and patterned with the stimulus-experiencer (67.6%) and presupposition verbs (61.5%). Fully consistent with our model, the explanations of experiencer-stimulus verbs were overwhelmingly of the *simple cause* type (87.9%) and led to a strong NP2 bias (14.7% NP1). Finally, even for verbs without underspecified content, i.e. agent-patient verbs without a presupposition, we were able to account for most of the variance in IC bias by correlating it with the ratio of external and internal reasons (linear regression: corrected $R^2 = .75$). Fig. 4 shows that the ratio of internal and external reasons is a good predictor of IC bias. Also in line with our predictions, the figure shows a gradient transition from 100% external reasons to 96% internal reasons which offers a straightforward explanation for the observed variability of IC bias in this class of verbs.

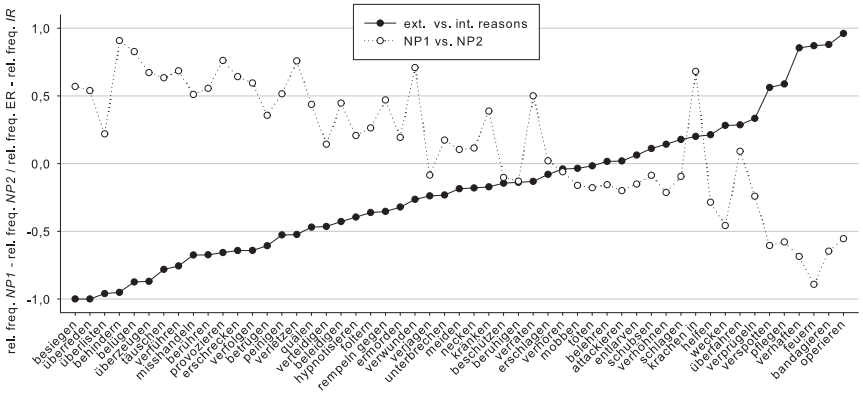


Figure 4: Unambiguous AP verbs in Exp. 2. IC bias (white dots) is shown as a function of the ratio of internal (IR) and external reasons (ER) (black dots). Bias: +1.0 corresponds to 100% NP1 and −1.0 to 100% N2 bias. Reasons: −1.0 corresponds to 100% IR, +1.0 to 100% ER.

Conclusions

The results show that IC bias strongly depends on the availability of specific explanation types and that it can be manipulated by specifying implicit explanations in the prompt. The proposed analysis offers a formally precise, predictive model that can account for a large body of processing data (off- and online) that have accumulated over the last four decades, having been largely neglected in semantic/pragmatic theory. At the same time, it yields fine-grained, novel predictions as to the incremental nature of explanatory discourse which are highly relevant for psycholinguists interested in the underlying forces of establishing discourse coherence.

References

[1] M. Aloni and F. Roelofsen. Interpreting concealed questions. *L&P*, 34:443–478, 2011.

[2] G. Altmann and M. Steedman. Interaction with context during human sentence processing. *Cognition*, 30:191–238, 1988.

[3] O. Bott and T. Solstad. From verbs to discourse: a novel account of implicit causality. In B. Hemforth, B. Schmiedtová, and C. Fabricius-Hansen, editors, *Psycholinguistic approaches to meaning and understanding across language*. Springer, to appear.

[4] R. Brown and D. Fish. The psychological causality implicit in language. *Cognition*, 14:237–273, 1983.

[5] E. C. Ferstl, A. Garnham, and C. Manouilidou. Implicit causality bias in English: A corpus of 300 verbs. *Behav Res Methods*, 43:124–135, 2011.

[6] J. K. Hartshorne and J. Snedeker. Verb argument structure predicts implicit causality: The advantages of finer-grained semantics. *Language and Cognitive Processes*, 28:1474–1508, 2013.

[7] A. Kehler, L. Kertz, H. Rohde, and J. L. Elman. Coherence and coreference revisited. *JoS*, 25:1–44, 2008.

[8] P. Pyykkönen and J. Järvikivi. Activation and persistence of implicit causality information in spoken language comprehension. *Exp Psychol*, 57:5–16, 2010.

[9] T. Solstad. Some new observations on 'because (of)'. In M. Aloni, H. Bastiaanse, T. de Jager, and K. Schulz, editors, *Amsterdam Colloquium 2009*, LNCS. Springer, 2010.

[10] R. A. van der Sandt. Presupposition projection as anaphora resolution. *JoS*, 9:333–377, 1992.