Only *Only* ? An Experimental Window on Exclusiveness*

Jacques Jayez[1] and Bob van Tiel[2]

[1] ENS de Lyon and lℒC$_2$, CNRS, Lyon, France
[2] Radboud University, Nijmegen, The Netherlands

Abstract. Beaver and Clark (2008) recently argued that *only* φ does not presuppose the proposition in its scope, contrary to the 'standard' theory articulated by Horn (1969). Their rejection of the standard theory is partially based on results of a survey test. We present new experimental evidence challenging Beaver and Clark's interpretation of this survey test and suggesting that dropping the standard theory altogether might be too radical a move.

1 Introduction

According to Horn's (1969) famous analysis, a sentence of the form *only* φ presupposes its prejacent φ and asserts that all alternatives to φ are false. For example, (1a) presupposes ($\rightsquigarrow$) that Paul smokes and asserts ($\Rightarrow$) that nobody else does. In support of this view, the prejacent appears to project from under sentential negation, as (2) suggests.

(1) Only Paul smokes.
$\rightsquigarrow$ Paul smokes.
$\Rightarrow$ Nobody else smokes.

(2) It is not the case that only Paul smokes.
$\rightsquigarrow$ Paul smokes.
$\not\Rightarrow$ Nobody else smokes.

Beaver and Clark (2008) (B&C) raise a number of objections against the standard theory and argue that *only* presupposes that the prejacent is the weakest proposition on a scale and asserts that it is the strongest proposition on the same scale. 'Weak' and 'strong' are not defined in terms of logical entailment, but in a more liberal way which is compatible with scales based on degrees of importance or cardinality.

In this paper, we reconsider the empirical and especially the experimental evidence laid out by B&C to back up their claim. The gist of B&C's experimental data is that the prejacent of *only* behaves differently from other presuppositions.

* We gratefully acknowledge the financial support of the European Science Foundation, ESF travel grant Euro-Xprag 4273.

M. Aloni et al. (Eds.): Amsterdam Colloquium 2011, LNCS 7218, pp. 391–400, 2012.

We present results of a more extensive experiment that challenge this conclusion on two counts. First, other presupposition triggers with an entirely different semantics behave in the same way as *only*. Second, a semantically very similar trigger (*seulement* in French) behaves quite differently. This runs counter to B&C's claim that *only* has a somehow special status compared to other presupposition triggers.

In sections 2.1, 2.2 and 2.3, we present B&C's objections against the standard theory and describe their alternative proposal. In section 2.4, we show that their empirical arguments are inconclusive. In section 3, we present the results of an extensive survey, which will be discussed in section 4. These results undermine B&C's critique of the standard theory.

2 Beaver and Clark's Approach

2.1 Linguistic Observations

According to the standard theory, *only* φ presupposes its prejacent. This means that the prejacent is predicted to project. Against this prediction, B&C observe that there are cases in which the prejacent is not preserved under negation. Consider (3). It is clear that the speaker does not imply that the person in question is a blond bimbo with no brains.

(3) She's one of the first that really represents the country and isn't only some blond bimbo with no brains.

Another piece of evidence in the same direction is provided by examples like (4), which does not entail that Mary invited Susan and Paul, since she invited their six cousins instead.

(4) Last year, Mary invited Susan and Paul. This year, she did not invite only Susan and Paul, but preferred to invite their six cousins.

These examples seem to endanger the standard theory. If *only* φ presupposes its prejacent, why does it not project in the examples above?

2.2 The Tequila Test

To further underpin the view that the prejacent of *only* is more 'fragile' than presuppositions of other triggers, B&C devised an experiment based on the interpretation of a little story:

One year there were 90 students in Arroyo.
30 drank Tequila and nothing else.
30 drank non-alcoholic beverages and nothing else.
30 drank everything, no matter what.

Subjects had to answer the following two questions: *How many students didn't only drink Tequila* (VP-*only*) and *How many students didn't drink only Tequila* (NP-*only*). They could choose between the following answers: '30', '60' and 'Don't know'.

The standard theory predicts most participants to opt for the '30' answer. After all, only the 30 students who drank everything satisfy both the prejacent (i.e., they drank Tequila) and the asserted content (i.e., they drank something else). If, on the contrary, the prejacent is not genuinely presupposed, participants might also include the 30 students who drank non-alcoholic beverages and nothing else in the denotation of the question. While these students do not satisfy the prejacent, they do fulfill the asserted content. In that case, participants should choose the '60' answer. Assuming that the first set is ruled out in any case, it might seem that the 'Tequila test' is an adequate instrument to decide whether the prejacent is presupposed.

B&C report the following results (absolute numbers between brackets):

	'30'	'60'	Don't know
NP-*only*	22% (9)	76% (31)	2.4% (1)
VP-*only*	41% (17)	56% (23)	2.4% (1)

There is a substantial difference between the results in the NP-*only* and VP-*only* condition. Because the subjects were not divided into two independent or paired samples, it is difficult to interpret these results in a reliable way. It is possible to run a McNemar's test on the results, under the assumption that the subjects are 'coherent'; that is, that the subjects who chose '30' for NP-*only* are a subset of those who chose '30' for VP-*only* and that the subjects who chose '60' for VP-*only* still chose '60' for NP-*only*. In that case, the difference between the two positions for *only* is significant at the 0.05 threshold (p-value ≈ 0.012).[1] That is, the answer '60' was chosen significantly more often in the NP-*only* condition than in the VP-*only* condition.

Regardless of this difference, there was overall a high number of '60' answers, particularly in the NP-*only* condition. These results appear to jeopardise the view that the prejacent is presupposed. But for this argument to go through, it still has to be shown that other presupposition triggers lead to significantly different results. To show this, B&C used a comparable testing procedure for four other presupposition triggers: *stop*, *realize*, *regret* and *their*. The set-up was basically the same. Participants read a short story followed by a question in which the trigger was embedded under negation. The story involved 90 students, 30 of which satisfied the presupposition and the asserted content, 30 of which satisfied only the asserted content, and 30 of which satisfied neither the presupposition nor the asserted content (or in some cases the presupposition but not the asserted

[1] B&C report a non-significant result for a chi-square test. The problem with using this test is twofold: if the subjects are coherent, in the sense considered here, the chi-square is not a good indicator. If they are not coherent, to a degree that falsifies our assumption, the question is more complex because the shift in perception that this incoherence suggests has to explained.

content).[2] The critical question was whether participants include the students who falsified the presupposition but not the asserted content in the denotation of the question. The target item for *stop* is provided below:

There are 90 students:
30 used to drink but gave up.
30 never drank Tequila.
30 currently drink Tequila.
How many students didn't stop drinking Tequila?

If the question presupposes that the students used to drink Tequila, the correct answer is '30'. If this presupposition is as 'fragile' as the prejacent for *only*, participants might opt for '60', thus including the students that never drank Tequila. The results for these presupposition triggers clearly differ from the results that were found for *only*. For *stop*, *realize* and *their*, 9, 12 and 10 participants out of a total of 13, chose the answer which is compatible with the projection of the presupposition, namely '30'. These findings seem to indicate that the prejacent has a somewhat different status than ordinary presuppositions. This compromises the standard theory of *only*, but fits in neatly with B&C's counterproposal, to which we now turn.

2.3 The Proposal

B&C propose to amend the standard theory by exploiting the scalar character of *only*. To put it concisely, *only* φ presupposes that the prejacent is at most as strong and asserts that it is at least as strong as any true alternative. More precisely, we have:

> *Only* p presupposes (asserts) that for every proposition q in an appropriate set of alternatives to p, $ALT(p)$, if q is true then p is at most (at least) as strong as q. In symbols: *only* p presupposes the propositions defined by $\lambda w \forall q \in ALT_\sigma(p)(w \models q \Rightarrow q \geq_\sigma p)$, and asserts the propositions defined by $\lambda w \forall q \in ALT_\sigma(p)(w \models q \Rightarrow q \leq_\sigma p)$, where σ is the belief state of the speaker.

Let us apply this definition to (1), assuming that the set of alternatives is calculated on the basis of alternatives of the form 'x smokes', which are ordered by entailment. Here, x ranges over a set of possible persons or groups. The presupposition eliminates worlds in which 'Paul smokes' is stronger than some alternative

[2] The reason for this variation is that it is very difficult for some triggers to describe a character that falsifies both the asserted content and the presupposition. For *Who didn't stop drinking?*, for example, this would amount to a character who at the very moment of uttering the question started drinking. Such a character falsifies both the presupposition (i.e., she didn't drink before) and the asserted content (i.e., she doesn't drink now). But it is hard to exclude the possibility that this character actually did drink at least some time before the moment of utterance, thus verifying the presupposition. To avoid this ambiguity, B&C construed a character that verifies the presupposition in these cases.

which is true at the same world. The common ground is then updated with the main content. This move eliminates worlds in which there is a proposition of the form 'x smokes' which is stronger than 'Paul smokes'. For instance it eliminates worlds in which Paul and Mary, or Paul and John smoke. The net result is a set of worlds where, for each true proposition $q \in ALT_\sigma(p)$, $q =_\sigma p$. If we apply a negation to *Only Paul smokes*, the presupposition is (normally) preserved but the main content is negated. So, the negated sentence asserts that $\lambda w \exists q \in ALT_\sigma(p)(w \models q \,\&\, q >_\sigma p)$, in other words, that Paul *and* someone else smoke.

When alternatives are not ordered by entailment, a different result can obtain. For instance, if a cardinality-based ordering is used, the presupposition is that the prejacent concerns at most as many individuals as any true alternative. This delivers the required reading for (4). The negated main content entails that Mary invited more persons than just two. However, it does *not* entail that the guests include Susan and Paul, since the alternatives are compared on the basis of cardinality and not of entailment.

Summarising, B&C (i) replace the prejacent with a lower bound on the relative strength of the true alternatives, and (ii) assume that the main content sets an upper bound on the strength of the true alternatives. The derivation of the prejacent is thus an effect of the interaction between these two constraints, not an intrinsic semantic property of *only*.

2.4 Preliminary Discussion

One might wonder why it is necessary at all to gather experimental data in order to decide between different theories of *only*. In examples like (4), the prejacent manifestly does not project. Doesn't this itself conclusively disprove the standard theory? In fact not. Examples like (4) can be construed for any presupposition trigger. For example, B's answer in (5) clearly does not imply that John has been smoking recently. Such examples can be explained as local accommodation or metalinguistic negation (e.g., Geurts 1999).

(5) [Context: it is common belief that John never smoked. B is trying to quit.]

A – John seems to be much more relaxed than you are.
B – He didn't stop smoking a week ago!

On the whole, B&C's empirical observations are not as conclusive as they may seem at a first glance because they are not restricted to *only*, but concern rather the pragmatic conditions on the felicity of presupposing. So, it is not clear that a specific theory should be constructed for *only* on the basis of examples such as (3) or (4).[3]

This leaves us with the results of the Tequila test. Although B&C's empirical observations do not support the view that the prejacent is different from an

[3] A similar remark applies to Ippolito's (2008, 50-52) discussion about *it's possible that only* φ. As shown by Herburger (2000, 95) the observations that would tend to show that the prejacent is suspended are not specific to *only*. B&C mention Herburger's work and add further examples which suggest that *only* is not the main factor in those cases (Beaver and Clark 2008, 245-246).

ordinary presupposition, the results of their Tequila test clearly do. Hence the importance of carefully evaluating the validity of B&C's experimental results.

3 An Experimental Approach

The Tequila test, as carried out by B&C, faces several more or less severe problems. First, one of their surveys included a very limited number of 13 participants. Second, no fillers were included in any of the surveys. This was especially pertinent for the experiment that involved *only*. The entire experiment consisted of two nearly identical questions, differing only in the position of *only* in the sentence (before the VP or just before the NP). It is quite possible that this juxtaposition led participants to explicitly contrast the two questions. As a final worry, there were some presentational and interpretational differences between the story involving *only* and the stories involving the other presupposition triggers.

A perhaps more interesting issue concerns the scope of B&C's Tequila test. First, the selection of presupposition triggers B&C employed to compare with *only* was rather limited. Second, it might be hypothesised that B&C's results are somehow peculiar to English *only*. This idea is fueled by our intuitions about the Dutch and French equivalents of *only*. These considerations led us to form the following hypotheses: (i) For English, there are presupposition triggers that behave just like *only*; (ii) The prejacent of the Dutch and French equivalents of *only* are not as 'fragile' as the prejacent of English *only*. The truth of either hypothesis would disprove B&C's conclusion that the prejacent of *only* has a special status compared with other presuppositions.

3.1 The Basic Protocol

In order to remedy some of the methodological issues of B&C's survey, we made some changes to the design of the Tequila test. Instead of using numbers, we used characters. Just like in B&C's version, one character satisfied neither the asserted content nor the presupposition of the question (or in some cases the presupposition but not the asserted content). One character verified the asserted content but falsified the presupposition. One character verified both the asserted content and the presupposition. These characters correspond to respectively A, B and C in the examples below. Again, the critical issue was whether participants include the B character in the denotation of the question.

We ran the experiment in three languages: Dutch, English and French. The target stimuli were interspersed with filler stimuli in the same vein but using various quantifiers such as *at most three* or *often*. The stimuli and the attribution of the actions or situations to A, B and C were pseudo-randomised. We had 16 presupposition triggers and 16 fillers for Dutch, the same numbers for English, and 15 triggers and 23 fillers for French. The triggers included focus particles like *only* or *also*, factives like *know* or *regret*, implicatives like *manage* or *succeed*, aspectuals like *stop* or *start* and definites like *the* or *all*. For English, participants were recruited through Amazon MTurk. For French and Dutch, university students were asked to fill out the experiment. After we got the results,

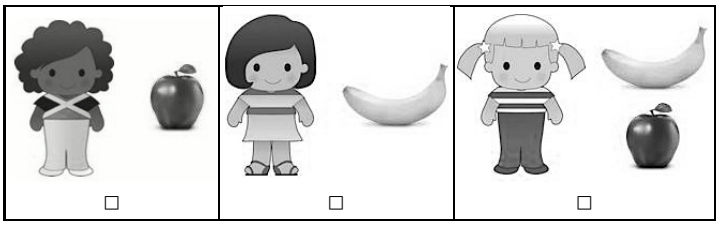
Only	Other triggers
Three people were in the cafeteria	Three people are riding a bus
A drank orange juice and nothing else	A had a job at the bank but quit
B drank coffee and nothing else	B never had a job in her life
C drank orange juice and coffee	C has a job at the bank and still works there
Who didn't drink only orange juice?	Who didn't resign from the bank?
□ C	□ C
□ C and B	□ C and B
□ I don't know	□ I don't know

Fig. 1. Two target stimuli

we decided to eliminate the *démissionner* ('resign') case from the French data, because the little story associated with it was problematic.

3.2 The Image-Based Protocol

When it turned out that the results for English *only* were markedly distinct from the results for its Dutch and French equivalents, as will be explained in the next section, we decided to run an additional experiment for English speakers, based on the expectation that the linguistic presentation of possible answers may have influenced participants. In this experiment, participants were presented with series of three images and had to answer a question that was completely analogous to the question asked in the Tequila paradigm. Participants could tick any number of boxes they liked. The critical stimulus is shown in figure 2.



Who does not have only an apple?

Fig. 2. An image-based target stimulus

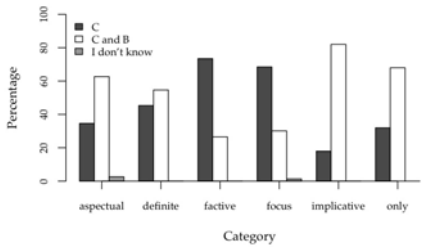
25 participants were drafted through Amazon MTurk. The experiment consisted of 6 stimuli, including 1 target stimulus featuring *only* and 5 fillers involving quantifiers. The stimuli were pseudo-randomised.

3.3 Results

The comparison between different kinds of English triggers is summarised graphically in figure 3. The left (black) column represents the percentage of 'C'

answers, the middle (white) column the percentage of 'B and C' answers and the right one (grey) the percentage of 'I don't know' answers. Because there are very few 'I don't know' answers, it is possible to binarize the results by dividing the answers into 'C' versus other answers ('B and C' and 'I don't know'). The dependent variable is the proportion of 'C' answers. The independent variables are language and type of stimulus, e.g. implicatives, factives, etc.

We analysed these data by means of a logistic regression analysis. Using the 'lme4' package in R, we fitted a simple model of mixed logistic regression, having the subjects as random effect and adding a post hoc comparison based on the 'multcomp' package. The results are summarised in figure 3 for English *only*. The number of 'C' responses for this item is compared to the number of 'C' responses for other kinds of triggers. The difference is significant for factive and focus elements and non-significant for other categories.



	z value	Pr(>|z|)
vs. aspectual	-0.261	0.99983
vs. definite	-1.249	0.80606
vs. factive	-3.892	0.00139**
vs. focus	-3.289	0.01233*
vs. implicative	1.419	0.70749

Fig. 3. English triggers

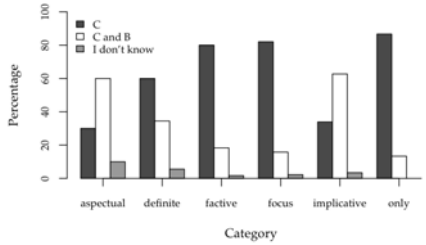
The pictorial task for English illustrated in figure 2 gave totally consonant results. For the critical item involving *only*, 72% of speakers chose the 'B and C' answer. So it seems unlikely that the linguistic nature of the test was an issue.

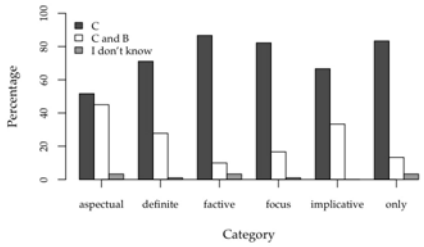
The Dutch and French counterparts of *only*, *alleen* and *seulement*, do not behave like *only* in English. The two relevant histograms and the post hoc contrasts are shown in figure 4. The post hoc contrasts on a simple logistic regression with the response binary variable restricted to the *only* case show that Dutch and French are not significantly different whereas they are both different from English (Pr(>|z|) = 0.93, 0.0004, 0.0008).

4 Discussion and Perspectives

The language-based and image-based results for English are consonant with B&C's observations for *only* and the other presupposition triggers they investigated. This shows that our results were not affected by the methodological changes we made to the Tequila paradigm.

Our overall results call into question B&C's conclusion that the prejacent of *only* behaves differently from ordinary presupposition triggers. Figure 3 shows no difference between *only* and aspectuals, implicatives, or definites. This is unexpected if the relation of *only* to its prejacent is specific. If we assume, with B&C, that the preponderance of 'C and B' answers for *only* suggests that the prejacent





DUTCH	z value	Pr(>|z|)
vs. aspectual	5.003	< 0.001 ***
vs. definite	2.704	0.06978.
vs. factive	0.869	0.95116
vs. focus	0.613	0.98948
vs. implicative	4.528	< 0.001***

FRENCH	z value	Pr(>|z|)
vs. aspectual	2.893	0.0422*
vs. definite	1.344	0.7523
vs. factive	-0.476	0.9968
vs. focus	0.141	1.0000
vs. implicative	1.516	0.6443

Fig. 4. Dutch and French triggers

is not genuinely presupposed, we would arrive at the counterintuitive conclusion that the presuppositions ordinarily associated with aspectuals, implicatives, and definites are not genuine presuppositions either. Whatever conclusions one might draw from the behavior of *only*, it is clear that these do not hold for *alleen* and *seulement*. For these triggers, participants almost unanimously opted for the 'C' answer. Adopting B&C's interpretation again leads to the implausible conclusion that the prejacent of *alleen* φ and *seulement* φ is genuinely presupposed but the prejacent of *only* φ is not.

One might wonder whether the observed profiles coincide with the 'weak' versus 'strong' trigger distinction made by Abusch (2010). Abusch contrasts examples like those in (6). *Win*, which presupposes a participation in the competition, allows for the suspension of its presupposition and is, in this respect, 'weak', in contrast to *again*, which is a 'strong' trigger.

(6) a. I don't know whether John finally participated in the race, but if he won it he may be very proud!
b. ?? I don't know whether John won this race before, but if he won again, he may be very proud!

At a first glance, it might seem that strong triggers evoke mostly 'C' answers whereas weak triggers evoke mostly 'C and B' answers. This indeed holds for the strong triggers *again* and *also*. Unfortunately, the parallelism breaks down when it comes to aspectuals, which are presumably weak. We found perhaps surprisingly that *start* leads to mostly 'C and B' answers whereas *stop* evokes mostly 'C' answers. Overall, the data do not correspond to a systematic weak/strong distinction.

An important issue in the semantics of exclusives is their scalar character. It is well-known that *only* is scalar in that it can be interpreted as entailing that the degrees above or below a certain threshold, as expressed by the prejacent, are not reached. French *seulement* has the same property (e.g., Beyssade 2010), whereas

alleen is not scalar, see (7). It is remarkable, then, that *seulement* patterns with *alleen* and not with *only*, with which it shares its scalar nature.

(7) a. Paul is only a first-year student.
b. Paul est seulement un étudiant de première année.
c. *Paul is alleen een eerstejaars student.

The reported observations raise a more general question. It is often (partly) implicitly assumed that the behaviour of presupposition triggers should be a reflection of their formal semantics, because assuming the contrary would lead us to renounce any explanation. In our opinion, this dilemma lacks serious foundations. The high cross-linguistic variability of certain triggers, which sound otherwise quite comparable, comes as a surprise under this view, but remains compatible with an approach that is not (entirely) representational, in which triggers *in addition* to a descriptive content (main content + presupposed content) have a statistical profile with respect to, say, suspension under negation or other environments. It remains to see what factors cause these differences in statistical profile.

In future work, we intend to tighten the experimental conditions, by having a homogeneous pool of subjects in the three languages and controlling the stimuli and the choice of answers even more carefully. We are also planning two new experiments, one using reaction times, in order to determine whether there is any correlation between the 'B and C' answer and the choice duration. We also intend to test whether the observed difference might be connected with the 'loneliness' flavour associated with *seulement* and *alleen*, both of which provide adjectives meaning 'alone', in contrast to English (**Paul is only*). To this aim, we will turn to languages similar to English in this respect like Chinese.

References

Abusch, D.: Presupposition triggering from alternatives. Journal of Semantics 27, 37–80 (2010)
Beaver, D., Clark, B.Z.: Sense and Sensitivity. How Focus Determines Meaning. Wiley-Blackwell, Chichester (2008)
Beyssade, C.: Seulement et ses usages scalaires. Langue Français 165, 103–124 (2010)
Geurts, B.: Presuppositions and Pronouns. Elsevier, Amsterdam (1999)
Herburger, E.: What Counts: Focus and Quantification. MIT Press, Cambridge (2000)
Horn, L.: A Presuppositional Analysis of only and even. Papers from the Fifth Regional Meeting of the Chicago Linguistics Society, pp. 98–107 (1969)
Ippolito, M.: On the Meaning of Only. Journal of Semantics 25, 45–91 (2008)
Roberts, C.: Only, presupposition and implicature. MS., Ohio State University (2006), `http://ling.osu.edu/~croberts/only.pdf`