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, Netherlands

Abstract. In a recent book, David Beaver and Brady Z. Clark argue that *only* does not presuppose the proposition in its scope, contrary to the 'standard' theory articulated in (Horn 1969), and propose a new semantic representation for *only*. Their rejection of the standard theory is based on linguistic observations and a survey test. We adduce new experimental evidence suggesting that (i) dropping the standard theory altogether might be too radical a move and (ii) the very idea that the behaviour of *only* derives from a combination logical form + context, even if it proves ultimately adequate, cannot be taken as granted.

1 Introduction

According to Horn's (1969) famous analysis, *Only* p, where p is a proposition, presupposes that its *prejacent*, p, is true and asserts that, for some salient set of alternatives to p, $Alt(p)$, every member of $Alt(p)$ different from p is false. So, (1a) presupposes that Paul smokes and asserts that no other person from a salient set does. (1b) shows that *only* behaves like a standard presupposition trigger. The presupposition that Paul smokes projects under sentential negation and interrogation.

(1) a. Only Paul smokes.
b. It is not true that only Paul smokes. / Does only Paul smoke?

Beaver and Clark (2008) (B&C) raise a number of objections against this approach 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 cardinals.

In this paper, we reconsider the empirical and experimental evidence laid out by B&C to back up their claim. We show that it is problematic in certain respects and that it does not support the view that the observations for *only* can be derived from the scalar semantics they propose, essentially because other presuppositional triggers with a quite different semantics behave in the same

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

way and a semantically very similar trigger (*seulement* in French) behaves quite differently.

In sections 2.1, 2.2 and 2.3, we present B&C's arguments and theory. In section 2.4, we show that their linguistic arguments are not conclusive. In section 3, we present the results of a new experimental approach and discuss them in section 4, showing that they undermine the very idea that *only* is a special item, whose particular behaviour is a reflection of its logical form.

2 Beaver and Clark's Approach

2.1 Linguistic Observations

B&C observe that there are cases in which the prejacent is not preserved under negation. Consider (2), their example 9.42.c. It is clear that the speaker does not believe that the person in question is a blond bimbo with no brains.

(2) 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 (3), which compares cardinals and does not entail that Mary invited Susan and Paul, since she invited their six cousins instead.

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

B&C note that such examples run counter to Horn's (1969) initial proposal and to Robert's (2006) defence of the standard theory as well as to Ippolito's (2008) implicative theory, which assumes that *if* some proposition is true in the set of alternatives associated with *Only p*, then *p* is true. Under that analysis, for instance, the prejacent of (2) is predicted to project, because the set of alternatives contains the proposition described by 'She is one of the first that really represents the country'.

2.2 The Tequila Test

In order to show 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* (VP-*only*). They had to pick one of the answers: '30', '60' and 'Don't know'.

Suppose that the subjects follow the standard theory. In that case, they should choose the '30' answer, since they would actually answer the following

question: 'For how many x is it true that x drank Tequila (the prejacent) and false that x didn't drink anything else?'. The third set is then the only correct answer. If, on the contrary, the prejacent is not preserved, subjects would answer the following question: 'For how many x is it true or false that x drank Tequila (the prejacent) and false that x didn't drink anything else?'. In that case the second and third sets are good candidates. Assuming that the first set is not a correct candidate, no matter if the prejacent is preserved or not, we see that the Tequila test might theoretically act as a separator between the two interpretations (with/without the prejacent).

B&C report that, out of 41 participants, 17 chose answer '30' for VP-*only* and 9 for NP-*only*, whereas 23 chose answer '60' for VP-*only* and 31 for NP-*only*. Because the subjects were not divided into two independent or paired samples, it is difficult to interpret these results in a reliable way, but they seem problematic for the standard theory. The number of '60' answers is particularly high for NP-*only*. 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).[3]

In addition to *only*, B&C used a comparable testing procedure for four other presupposition triggers: *stop*, *realize*, *regret* and *their*. The results are clearly different than for *only*. For instance, with *stop*, *realize* and *regret*, 9, 12 and 10 subjects out of a total of 13, choose the answer that is compatible with the projection of the presupposition. These results might be taken to suggest that *only* is special, at least as regards the projection of the prejacent under negation.

2.3 The Proposal

B&C propose to amend the standard theory by exploiting the scalar character of *only*. *Only* 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 (4).

(4) *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 proposition 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 see what happens with (1a), assuming that the set of alternatives is calculated on the basis of a form 'x smokes', where x ranges over a set of possible

[3] B&C report a non-significant result for a chi-square test. The problem 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.

persons or groups, and that they are ordered with respect to entailment. The presupposition eliminates worlds in which 'Paul smokes' is stronger than some alternative 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 '*a* 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 the proposition corresponding to $\lambda w \exists q \in ALT_\sigma(p)(w \models q \,\&\, q >_\sigma p)$, in other terms, the proposition that Paul *and* someone else smoke.

When alternatives are not ordered with respect to entailment, a different result can obtain. For instance, if a cardinality-based ordering is used, the presupposition is the proposition that the prejacent concerns at most as many individuals as any true alternative. This delivers the required reading for (3). 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's approach consists in (i) replacing the prejacent with a condition on the relative strength of the prejacent and its competitor and (ii) using the opposite condition as the main content. The derivation of the prejacent is an effect of the interaction between the two constraints, not an intrinsic semantic property of *only*.

2.4 Preliminary Discussion

Two points are in need of clarification. First, if the ordering is based on cardinality, what happens with a sentence like *Mary invited only Susan and Paul*? Intuition says that Mary invited Susan and Paul, but this is not a direct consequence of the formal part of the theory, because it is compatible with a situation where Mary invited two persons, (partly) different from Susan and Paul. In that case, every true alternative is equivalent to 'Mary invited Susan and Paul'. This possibility is excluded because it would be totally misleading. Suppose a speaker wants to address the question *How many persons did Mary invite?* by conveying two pieces of information, (1) that Mary invited only two persons and (2) that they were John and Sandy. If that speaker mentioned Susan and Paul as guests, she would just sound incoherent, because nobody would just ignore the names *Susan* and *Paul* in order to get the intended message 'only two'. See the contrast in (5).

(5) A – How many persons did Mary invite?
B1 –Only two persons. John and Sandy.
B2 –?? Only Susan and Paul. John and Sandy.

But how is it that (3) is unproblematic? An obvious answer is that it is relevant to mention Susan and Paul because they were invited last year and there is a

contrast with the current situation. In the absence of a salient contrast, (3) is infelicitous as an answer to a *how many* question.

(6) [Context: B does not know that Mary had previously invited Susan and Paul]
A – How many persons did Mary invite?
B – # She invited six persons. She didn't invite only Susan and Paul.

Examples like (3) are taken into account in any theory of presupposition and thus do not seem to be specific to *only*. Suppose that B has been a chain-smoker for years and has suddenly decided to quit. The beginnings are difficult and, when A meets B, B is particularly nervous and tired. One year after, A meets B again and is surprised at how B looks better. In a dialogue like (7), B does not presuppose that he has been smoking recently. A similar observation holds for (2), see (8). Both examples are based on a contrast between different time periods or individuals.

(7) A – I'm glad to see that you are much better than last time!
B – Well, unlike last year, I'm not quitting smoking. Fortunately, it's behind me now! It has been six months since I have stopped smoking completely.

(8) [Context: it is common belief that John never smoked. B is trying to quit.]
A – John seems to be much more relax 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 one may wish because they do not seem to be 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 (2) or (3).[4] Still, the results of the Tequila test separate *only* from other presupposition triggers. We now turn to these experimental data.

3 An Experimental Approach

The Tequila test of B&C is affected by two problems. First, it does not satisfy some usual requirements for a survey test, e.g. on the number of subjects, the independence or pairing of samples and the presence of fillers. Second, it is based on the calculation of the size of certain subsets, a fact which might have an effect on the answers. If we have a set of people drinking only Tequila, the complement of this set includes those people who drank Tequila and something else *and* those people who drank something different from Tequila. It is unclear whether subjects interpreted the question *How many students didn't (only) drink (only) Tequila?* as bearing on the adverb or as a complementation question.

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

3.1 The Basic Protocol

In order to remedy some of these problems, we changed the design of the Tequila test. Instead of using numbers, we used characters, who are differentiated by their actions or situation. Typically a subject had to go through sequences like the following and tick one of the boxes.

Fig. 1. Two target stimuli

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

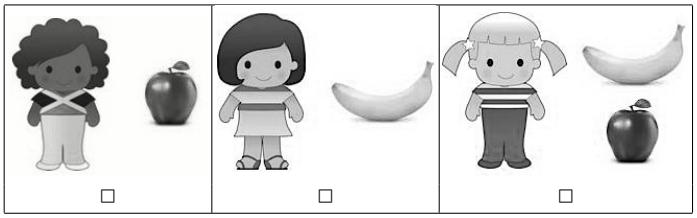
We ran experiments in three languages: Dutch, English and French. The target stimuli were interspersed with filler stimuli in the same vein but using various quantifiers such as *at most three* or *often*. The stimuli and the attribution of the actions/situations to A, B and C were pseudo-randomised. 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 include 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, subjects were recruited through Amazon MTurk. They were university students for Dutch and French. After we got the results, we decided to eliminate the *démissionner* ('resign') case from the French data, because the French little story associated with it was possibly problematic.

3.2 The image-based Protocol

When it turned out that the results for English were markedly distinct from those for Dutch and French, as explained in the next section, we decided to run an additional experiment for English speakers. Subjects were presented with series of three images and had to answer the same sort of question as for the test based on purely linguistic stimuli. An example of stimulus is shown in Figure 2.

There were 25 subjects who had to check 6 triples of images, including 1 target stimulus featuring *only* and 5 fillers again using quantifiers. The stimuli were pseudo-randomised as in the other experiments.

Fig. 2. An image-based target stimulus



Who does not have only an apple?

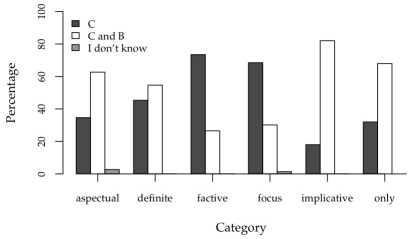
3.3 Results

In a nutshell, the observations collected for English support B&C's approach but the observations for Dutch and French do not.

The comparison between English triggers is summarised graphically in figure 3. The left (black) column represents the percentage of 'B' answers in figure 1, the middle (white) column the percentage of 'B and C' answers and the right one (grey) the '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 'B'-type versus others ('B and C' and 'I don't know'). The dependent variable is then the proportion of 'B' answers with respect to languages and types of stimuli, e.g. implicatives, factives, etc. There are several ways to analyse such data. One may run a McNemar test on every pair of stimuli and have a fine-grained image of the comparative distributions of answer. One may also use logistic regression, since the response is binary. Finally, another option is to cluster the items with respect to the binary response.

We first illustrate the case of logistic regression. 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 response for this item is compared to the overall response for different classes. The difference is significant for factive and focus elements and non-significant for other categories.

Fig. 3. English triggers



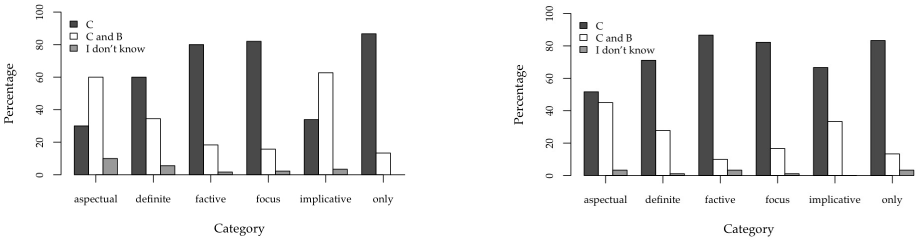
	z value	Pr(>|z|)
vs. aspectual	-0.261	0.99983
vs. definite	-1.249	0.80606
vs. factive	-3.892	0.00139**
vs. focus	-3.289	0.01233*
vs. implicative	1.419	0.70749

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

For Dutch and French, the counterparts of *only*, *alleen* and *seulement*, do not behave like in English. The two relevant histograms and the post hoc contrasts are shown in figure 4. In addition, 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 ($Pr(>|z|) = 0.93, 0.0004, 0.0008$).

We see that Dutch and French are similar except for implicatives. Since we had a difference for implicative stimuli between the two languages[5], we fitted two other separate models for the two cases of Dutch ('succeed' and 'manage'). The difference with French turned out to stable. Another, less marked, difference concerns the definite article. This difference cannot be reliably taken into account, because of a difference in the structure of the stimuli ('a professor' in Dutch versus 'one of his professors' in French).

Fig. 4. Dutch and French triggers



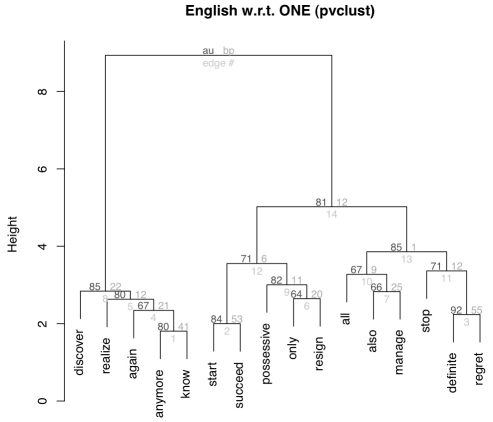
DUTCH	z value	Pr(>|z|)
vs. aspectual	5.003	< 0.001 ***
vs. definite	2.704	0.06978.
vs. factive	0.869	0.95116
vs. focus	0.613	0.98948
vs. implicative	4.528	< 0.001***

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

Probabilistic clustering with the pvclust package allows one to sharpen the general conclusions provided by using the mixed models, see figure 5. *Only* was found different from factives in the mixed model. Similarly, the 'regret', 'know', 'discover' and 'realize' nodes are far apart from the 'only' node in the cluster. This is true to a lesser extent for 'anymore', 'also', 'again', which form the focus category minus 'only'. Similar correspondences hold for Dutch and French clusters.

[5] This is due to the fact that, in French, *réussir à* ('succeed', 'manage') contains the same verb as *réussir* + NP ('pass', an exam for instance) and that *succeed* and *manage* are translated by *réussir*.

Fig. 5. Clustering for English



4 Discussion and perspectives

The language-based and image-based results for English are consonant with B&C's observations for the part of the Tequila test that concerns *only*. However, we did not find the same results for other triggers. Figure 3 shows no difference between *only* and the case of aspectuals, implicatives, or definites. This is unexpected if the relation of *only* to its prejacent is specific. Moreover, if the fact that the prejacent is suspended under negation is taken as an indication that *only* does not presuppose its prejacent, it seems that we cannot escape the same conclusion for the triggers that pattern with *only*.

One might wonder whether the observed profiles coincide with the 'weak' versus 'strong' trigger distinction in (Abusch 2010). Abusch contrasts examples like those in (9). *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

(9) 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!

Again is one of the strong focal elements compared with *only* and it is indeed clearly separated from it in the clustering (figure 5). This extends to *also*. Unfortunately, the parallelism breaks down when it comes to aspectuals and factives, which are presumably weak. *Start* is akin to *only* whereas *stop* is in a distinct subgroup. *Discover*, *realize* and *know* are also separated from *only*. In Dutch, *again* is close to *only*, as is *anymore* in French. 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 at least in that it can be interpreted as

entailing that the degrees above or below a certain threshold, expressed by the prejacent, are not reached. The French *seulement* has the same property, see (Beyssade 2010), whereas *alleen* is not scalar, see (10). However, both items resist suspension of the prejacent.

(10) a. Paul is only a first-year student (⇒ he is not a second/third/dots-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 distribution of 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 between calculation from a logical form and sterility lacks serious foundations. The high cross-linguistic variability of certain triggers, which sound otherwise quite comparable, comes as some surprise under this view, but remains compatible with an approach that is not (entirely) representational, where 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. This profile does not necessarily *derive* from something else.It could be an 'intrinsic' property of the item, that is, a property stabilised after some learning.

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*, which both 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 (e.g. Chinese).

References

Abusch, D. (2010). Presupposition triggering from alternatives. *Journal of Semantics* 27, 37-80.

Beaver, D. and Clark, B.Z. (2008). *Sense and Sensitivity. How Focus Determines Meaning*. Wiley-Blackwell, Chichester.

Beyssade, C. (2010). *Seulement* et ses usages scalaires. *Langue Française* 165, 103-124.

Herburger, E. (2000). *What Counts: Focus and Quantification*. Cambridge (MA): MIT Press.

Horn, L. (1969).A Presuppositional Analysis of *only* and *even*. In *Papers from the Fifth Regional Meeting of the Chicago Linguistics Society*, 98–107.

Ippolito, M. (2008). On the Meaning of *Only*. *Journal of Semantics* 25, 45-91.

Roberts, C. (2006). *Only*, presupposition and implicature. MS., Ohio State University, `http://ling.osu.edu/~croberts/only.pdf`.