Donkey Anaphora in Non-Monotonic Environments*

Milica Denić
Universiteit van Amsterdam

Yasutada Sudo
University College London

September 13, 2021

1 Introduction

Donkey anaphora is a type of pronominal anaphora involving a pronoun that is semantically bound by a non-c-commanding indefinite in a quantificational context as in (1).[1]

(1) Every farmer who owns a donkey beats it.

Donkey anaphora is known to have two types of readings: the ∀-reading, according to which the donkey pronoun is interpreted as a universal quantifier, and the ∃-reading, according to which the donkey pronoun is interpreted as an existential quantifier. For instance, the ∀-reading of (1) corresponds to (2a), while the ∃-reading of (1) corresponds to (2b) [Brasoveanu, 2007, Champollion, Bumford, and Henderson, 2019, Chierchia, 1995, Cooper, 1979, Foppolo, 2008, Geurts, 2002, Kanazawa, 1994, Krifka, 1996, Rooth, 1987, Schubert and Pelletier, 1989, Sun, Rothschild, and Breheny, 2020, Yoon, 1994, 1996].[2]

(2) a. Every farmer who owns a donkey or donkeys beats *all* of their donkeys.
b. Every farmer who owns a donkey or donkeys beats *at least one* of their donkeys.

Sentences with donkey anaphora typically strongly favor one of the two interpretations; the less preferred interpretation typically needs to be made salient by the context

*We wish to thank Emmanuel Chemla for many important discussions and suggestions. We also thank Lucas Champollion, Keny Chatain, Gennaro Chierchia, Floris Roelofson, Benjamin Spector and the audiences of LINGUAE seminar at ENS, XPRAG 2019 and Amsterdam Colloquium 2019 for helpful feedback. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007- 2013) / ERC Grant Agreement n. 313610 and n. STG 716230 CoSaQ, and was supported by ANR-17-EURE-0017.

[1]We will focus on *quantificational donkeys* like (1), and do not discuss *conditional donkeys* like (i) in this paper. This is because it is easier to manipulate the quantificational force with a nominal quantifier.

(i) If Mary owns a donkey, she beats it.

[2]For some speakers (1) is associated with a uniqueness presupposition that every farmer who owns a donkey owns only one, and with this presupposition the ∃ and ∀ readings collapse to one reading. Our experimental results presented below indicate that this is not obligatory for our participants, and we will not discuss this potential complication any further. See Heim [1982], Rooth [1987], Kadmon [1987], Heim [1990], Kadmon [1990], Krifka [1996], Chierchia [1995], Champollion et al. [2019] for related discussion.

in order to surface. Which quantifier heads the sentence affects which reading of a donkey pronoun is more prominent (Chierchia 1995, Kanazawa 1994, Yoon 1994, 1996, Champollion et al. 2019). Generally, donkey pronouns in the scope of universal quantifiers (e.g. *every*, *all*, *each*, etc.) tend to receive ∀-readings. For instance, (2) is most naturally interpreted with a ∀-reading, (2a), but as Chierchia [1995] points out, when a context like (3) is provided, a ∃-reading, (2b), becomes available.

(3) The farmers of Ithaca, NY, are stressed out. They fight constantly with each other. Eventually, they decide to go to the local psycho-therapist. Her recommendation is that every farmer who has a donkey should beat it, and channel his/her aggressiveness in a way which, while still morally questionable, is arguably less dangerous from a social point of view. The farmers of Ithaca follow this recommendation and things indeed improve. [Chierchia, 1995, 64]

Similarly, ∃-readings of donkey anaphora in universally quantified sentences are also observed with examples that make the ∀-reading implausible to be true due to world knowledge, as in (4) [Chierchia, 1995, Kanazawa, 1994, Schubert and Pelletier, 1989].[3]

(4) a. Every person who has a dime will put it in the meter. [Chierchia, 1995, p. 63]
b. Everyone who has a donkey must donate its services for one day during the festival. [Schubert and Pelletier, 1989, p. 200]

Contrary to donkey pronouns in the scope of universal quantifiers, donkey pronouns in the scope of *no* and existential quantifiers like *a* and *some* preferentially receive ∃-readings.

(5) No farmer who owns a donkey beats it.
≈ No farmer who owns a donkey or donkeys beats *any* of their donkeys.

(6) A farmer who owns a donkey beats it.
≈ A farmer who owns a donkey or donkeys beats *at least one* of their donkeys.

Importantly, it is not the case that donkey pronouns in the scope of *no* and existential quantifiers cannot receive ∀-readings [Kanazawa, 1994, Chierchia, 1995, Geurts, 2002, Champollion et al., 2019]. In line with ∃-readings with universal quantifiers, ∀-readings with *no* and existential quantifiers can surface when the world knowledge makes the ∃-readings implausible. For instance, Chierchia [1995, p. 65] points out that (7) is most naturally read under the ∀-reading (see Yoon 1996, p. 229 for more examples).[4]

(7) No one who has an umbrella leaves it home on a day like this.

Similarly, (8) with an existential quantifier can receive a ∀-reading (see also Sun et al. 2020 for *more than two*).

(8) Some people that have an umbrella left it home today, although it was clear that it would rain.

[3] Chierchia [1995] attributes (4a) to Schubert and Pelletier [1989], but their original example on p. 200 is a conditional donkey, *If I have a quarter in my pocket, I'll put it in the parking meter.*

[4] Chierchia [1995] cites a manuscript version of Kanazawa [1994] for this example, but its published version does not contain it.

Thus, across sentences and contexts, one may find both ∀- and ∃-readings for each of the above quantifiers, but the overwhelming tendency across sentences and contexts is that ∀-readings are more commonly observed with quantifiers like *every* than with quantifiers like *no* and *some*, which seem to more commonly receive ∃-readings. The less preferred readings typically surface when the more preferred readings are made implausible by world knowledge. These patterns have been corroborated by experimental studies [Yoon, 1994, Geurts, 2002, Foppolo, 2008, Sun et al., 2020]. These empirical facts have led some researchers to postulate (at least) two classes of quantifiers with respect to which of the ∀- and ∃-construals they more commonly receive across sentences and contexts [Kanazawa, 1994, Geurts, 2002, Champollion et al., 2019]. We will be largely following this literature and talk about the 'default readings' *of these quantifiers*, as our results are evaluated against their theories.[5]

In particular, Kanazawa [1994] proposes the following generalization.

(9) a. Default readings for ↑MON↑[6] quantifiers (e.g. *a, some*) and ↓MON↓ quantifiers (e.g. *no*) are ∃-readings.
b. Default readings for ↑MON↓ quantifiers (e.g. *all*) and ↓MON↑ quantifiers (e.g. *not all*) are ∀-readings.

Compared to donkey anaphora involving monotonic quantifiers, the behavior of donkey pronouns in the scope of non-monotonic quantifiers is less well understood.[7] However, there are two studies that make explicit claims about them. On the one hand, Kanazawa [1994] discusses 'existential non-monotonic quantifiers' such as *exactly three*, which are non-monotonic with respect to both arguments, and remarks that donkey anaphora under them prefers ∃-readings, as in (10).

(10) Exactly three farmers that own a donkey beat it.

Kanazawa [1994] offers some conjectures about why this might be so, which we will discuss in some detail in Section 2. More recently, Champollion et al. [2019] put forward an alternative theory to Kanazawa's (1994) according to which the conditions for sentences like (10) to be semantically true amount to a 'conjunctive reading', which is the conjunction of the ∃- and ∀-readings. Specifically, according to their trivalent semantics, (10) is semantically true if and only if there are exactly three donkey-owning farmers who beat their donkeys, and they all beat all of their donkeys, is false if and only if the sentence is false under both ∃- and ∀-readings, and receives the third truth-value (i.e. it is neither true nor false) in situations in which either only the ∃-reading or only the ∀-reading is true. Champollion et al. furthermore allow for the possibility that certain

[5]To be clear, Champollion et al.'s aim is broader than just accounting for the default readings. In particular, their theory is equipped with a pragmatic component that gives the theory enough flexibility to derive a reading of a donkey sentence on a context-by-context basis, but as a consequence one can only talk about the reading of a donkey sentence with respect to some specific context (see the discussion immediately below). Yet, as Champollion et al. [see 2019, p. 24] themselves suggest, one could still speak of default readings under this view as readings that result from the semantics put in some default context (e.g. a 'fact finding context', as Champollion et al. suggest), or alternatively, from interactions between the semantics and speakers' (presumably probabilistic) intuitions about conversational context where these sentences are likely to be used. We will review their theory in greater detail in Section 6.

[6]Upward/downward arrow to the left/right of MON stands for upward/downward monotonicity of the left/right argument of the quantifier.

[7]The literature sometimes mentions proportional quantifiers like *most*, which are non-monotonic with respect to the NP argument.

situations where the sentence semantically receives the third truth-value are deemed to be practically indistinguishable from situations where the sentence is semantically true or false, depending on what the pragmatic context is like. In other words, people may report a donkey sentence to be true or false when the sentence semantically receives the third value. Since what can actually be observed is speakers' 'pragmatic responses', that is, the truth value judgment in a pragmatic context, rather than the semantics itself, the exact predictions of this theory with respect to how donkey sentences such as (10) will be interpreted can only be identified in reference to pragmatic factors that affect the reported truth value in situations where the sentence semantically receives the third truth-value. Champollion et al. are not entirely explicit on how sentences with non-trivially trivalent semantics like (10) are to be used, including in what contexts they can and cannot be used felicitously. However, their account in theory allows for contexts where only the semantically true cases of (10) are considered pragmatically true, which is to say that (10) is understood to have a conjunctive interpretation in such contexts. We will discuss their account in light of our experimental results in Section 6.

To our knowledge, no controlled empirical study so far has compared donkey anaphora in different non-monotonic environments. The aim of this paper is to fill in this gap with an experimental study that compares donkey anaphora in the scope of two non-monotonic quantifiers, *exactly three* and *all but one*, using a truth value judgment task. We will explain why these two quantifiers were chosen in the next section.[8]

To preview the results of our experiments, we find that the ∃-reading is the most prominent one with both non-monotonic quantifiers. Interestingly, we find that donkey anaphora involving *all but one* receives ∀-readings more prominently than donkey anaphora involving *exactly three*, for which we see no evidence for a ∀-reading. This difference is, to our knowledge, not predicted by any existing approach to donkey anaphora. We also do not find conclusive evidence that the aforementioned conjunctive reading predicted by Champollion et al. [2019] is accessed for either of the two quantifiers.

The paper is structured as follows. We will first point out in the next section the theoretical possibility that some other logical property than monotonicity matters for preferred readings of donkey anaphora. This discussion will motivate the choice of non-monotonic quantifiers tested in our experiments. We will then present our main experiments in Sections 3 and 4. In Section 5 we present the results of an additional experiment investigating the interpretation of donkey anaphora with the quantifier *all*, which provides a helpful comparison point for our main experiments. In Section 6, we will discuss theoretical implications of our experimental findings with respect to Kanazawa's (1994) and Champollion et al.'s (2019) proposals. Section 7 contains conclusions and further directions. The experimental data, the R script used for analysis, and the design files can be found at https://github.com/milicaden/donkey-anaphora-nm.

2 Monotonicity, Symmetry and Left-Continuity

It is often considered in the literature that monotonicity is a major factor determining the default readings of quantifiers in donkey sentences. In particular, Kanazawa [1994]

[8]To be clear, our main interest in this paper is the overall truth-conditional intuitions of donkey sentences involving different quantifiers, and we don't have much to add to the debate about the compositional semantics of donkey anaphora. This is an aspect that is well discussed in previous work, e.g. Heim [1990], Kanazawa [1994], Chierchia [1995], Muskens [1996], Brasoveanu [2007], Champollion et al. [2019], among others, but remains controversial.

summarizes his generalization in terms of monotonicity, as we saw in (9) above, and proposes that this generalization can be explained in terms of *preservation of monotonicity*. The idea is that the reading of a donkey sentence that preserves the monotonicity of the quantifier in non-donkey sentences is the default reading. For example, the generalized quantifier corresponding to *no* is $\downarrow\mathrm{MON}\downarrow$, and in a donkey sentence, the $\exists$-reading, but not the $\forall$-reading, preserves this monotonicity profile with donkey anaphora. To see this, consider (11).

(11) a. No farmer who owns a donkey beats it.
b. No farmer who owns a young donkey beats it.

If both of these sentences received $\forall$-readings, then (11a) would not entail (11b). That is, (11a) under the $\forall$-reading would be compatible with a situation containing a farmer who owns young donkeys and beats all of them, but also has at least one old donkey that he doesn't beat. In this situation, the $\forall$-reading of (11a) would be true and the $\forall$-reading of (11b) would be false. On the other hand, the entailment from (11a) to (11b) would go through under the $\exists$-readings of the sentences. Consequently, under Kanazawa's (1994) proposal, the default reading of *no* in donkey sentences is the $\exists$-reading.

Let us look at another example, this time with a universal quantifier, which is $\downarrow\mathrm{MON}\uparrow$.

(12) a. Every farmer who owns a donkey beats it.
b. Every farmer who owns a young donkey beats it.

First consider the $\exists$-readings of these sentences, under which (12a) would not entail (12b). Specifically, in the following situation the $\exists$-reading of (12a) is true while the $\exists$-reading of (12b) is false: every donkey-owning farmer beats at least one old donkey he owns, but some of them don't beat any of the young ones. On the other hand, the entailment from (12a) to (12b) would go through if both sentences received $\forall$-readings. Therefore, according to Kanazawa [1994], the default reading of *every* is the $\forall$-reading.

This idea, however, cannot apply to non-monotonic quantifiers, simply because they have no monotonicity to preserve (or more precisely, their non-monotonicity is preserved under either reading of donkey anaphora). Yet, as Kanazawa [1994] points out, non-monotonic quantifiers like *exactly three* seem to prefer the $\exists$-reading. In order to make sense of this, Kanazawa [1994] suggests that this is because some other logical property, or properties, also need to be preserved, in addition to monotonicity. Specifically he conjectures that at least one of the following two logical properties might be the culprit: (i) *left-continuity* and (ii) *symmetry*.

For classical generalized quantifiers, these two properties are defined as follows, with D being the domain of the model.

(13) A quantifier Q is *left-continuous* if and only if for all $A, B, C, X \subseteq D$ such that $A \subseteq B \subseteq C$, $Q(A)(X)$ and $Q(C)(X)$ together imply $Q(B)(X)$.

(14) A quantifier Q is *symmetric* if and only if for all $A, B \subseteq D$, $Q(A)(B) = Q(B)(A)$.

The classical generalized quantifier corresponding to *exactly three* is both left-continuous and symmetric.[9] As Kanazawa [1994] observes, in donkey anaphora, the $\exists$-reading pre-

[9]The classical generalized quantifier corresponding to *exactly three* is: $[\![\text{exactly three}]\!](A)(B)$ iff $|A \cap B| = 3$. This is obviously symmetric. It is also left-continuous:

Proof. Suppose that $A \subseteq C$ and $|A \cap X| = |C \cap X| = 3$. Then for any B, if, $A \subseteq B$, then $|B \cap X| \geqslant 3$,

serves both of these properties, but not the ∀-reading. Let us see this first for the case of left-continuity. Consider the following sentences.

(15) a. Exactly three farmers that own an animal beat it.
b. Exactly three farmers that own a donkey beat it.
c. Exactly three farmers that own a young donkey beat it.

Take the following situation, which makes (15a) and (15c) true under the ∀-reading: Three farmers each own a young donkey and a cow (and nothing else), and beat both of them. Another farmer owns an old donkey and a cow (and nothing else), and only beats the donkey. There is no other farmer. Then (15b) is false, because there are four farmers who beat all of their donkeys. Therefore, left-continuity is not preserved under the ∀-reading. On the other hand, under the ∃-reading, whenever (15a) and (15c) are true, (15b) is also true because (15a) and (15c) together ensure that only young donkeys get beaten and their owners are the only farmers that beat any animals.

Let us now turn to symmetry. One issue is that symmetry under donkey anaphora cannot be checked by switching the two arguments of the quantifier, because that would disrupt anaphora. Fortunately, it is known that a conservative quantifier is symmetric iff it is intersective (e.g. Peters and Westerstål 2006), i.e. for all $A, B \subseteq D$, $Q(A)(B) = Q(A \cap B)(D)$, which can be checked without disrupting donkey anaphora.[10] Now, we observe that intersectivity is preserved under the ∃-reading, but not under the ∀-reading. For instance, consider the following sentences.

(16) a. Exactly three farmers that own a donkey beat it.
b. There are exactly three farmers that own a donkey that they beat.

Clearly, the ∃-reading of (16a) is equivalent to (16b) but its ∀-reading is not. Since we only discuss conservative quantifiers, we will not distinguish symmetry and intersectivity below.

Kanazawa [1994] proposes that if at least either of left-continuity or symmetry needs to be preserved, then the preference for ∃-readings for quantifiers like *exactly three* could be accounted for, but he does not provide a definitive answer as to whether both of them matter or not, and if only one of them does, which one. Thus, Kanazawa's hypothesis has the following three variants, depending on which properties need to be preserved.

1. symmetry + monotonicity
2. left-continuity + monotonicity
3. symmetry + left-continuity + monotonicity

and if $B \subseteq C$, then $|B \cap X| \leqslant 3$. Thus if $A \subseteq B \subseteq C$, then $|B \cap X| = 3$. □

[10] This is proved as follows:

Proof. Suppose Q is conservative and symmetric. Then by conservativity, $Q(A)(B)$ and $Q(A)(A \cap B)$ are equivalent. By symmetry, $Q(A \cap B)(A)$ is also equivalent, which in turn is equivalent to $Q(A \cap B)(A \cap B)$ by conservativity. Conservativity further implies that this is equivalent to $Q(A \cap B)(D_e)$. Now suppose that Q is conservative and intersective. Then $Q(A)(B)$ and $Q(A \cap B)(D_e)$ are equivalent. By conservatity, $Q(A \cap B)(A \cap B)$ is also equivalent, which is obviously symmetric. □

It should be remarked that the second and third variants can be simplified. The key observation is that all (left) monotonic quantifiers are left-continuous.[11] This means that for monotonic quantifiers, monotonicity preservation and left-continuity preservation make the same predictions. Therefore, if left-continuity needs to be preserved, there will be no independent evidence that monotonicity also needs to be preserved [pace Kanazawa, 1994].

In what follows we will report on experiments that are designed to tease apart the predictions of these variants of Kanazawa's (1994) hypothesis, by testing the default readings of two non-monotonic quantifiers, *exactly three* and *all but one*. Let us explain why we chose these two quantifiers.

Firstly, as Kanazawa points out, all these hypotheses predict that the default reading of *exactly three* is the $\exists$-reading. While this seems to be intuitively the case, we would like to obtain experimental corroboration of it. If it turns out that its default reading is the $\forall$-reading, then all the hypotheses need to be revised.

Secondly, the main difference between the first hypothesis and the others has to do with non-monotonic but non-symmetric and left-continuous quantifiers such as *all but one*.[12] Specifically, according to the first hypothesis, *all but one* has no property to preserve. This predicts that neither reading should be preferred. According to the other two hypotheses, the left-continuity of *all but one* needs to be preserved. This predicts that the reading which preserves its left-continuity will be preferred.

We saw above that the left-continuity of *exactly three* is preserved under the $\exists$-reading, but not under the $\forall$-reading. By contrast, the left-continuity of *all but one* is only preserved under the $\forall$-reading. In order to see this, consider the sentences in (17).

(17) a. All but one of the farmers who own an animal beat it.
b. All but one of the farmers who own a donkey beat it.
c. All but one of the farmers who own a young donkey beat it.

The following situation makes the $\exists$-readings of (17a) and (17c) true, but the $\exists$-reading of (17b) false: Farmer A owns some old donkeys and doesn't beat any of his animals; All the other farmers beat at least some of their animals; Farmer B owns some young donkeys, which he never beats, and horses, which he beats; All the other farmers who own young donkeys beat some of them. On the other hand, the $\forall$-readings of (17a) and

[11] This is proved as follows.

Proof. If Q is a $\uparrow$MON quantifier, then $Q(A)(X)$ entails $Q(A')(X)$ for any $A \subseteq A'$, so $Q(A)(X)$ and $Q(C)(X)$ together entail $Q(B)(X)$ for any B, C such that $A \subseteq B \subseteq C$. Similarly, if Q is a $\downarrow$MON quantifier, then $Q(A')(X)$ entails $Q(A)(X)$ for any $A \subseteq A'$, so $Q(A)(X)$ and $Q(C)(X)$ together entail $Q(B)(X)$ for any B, C such that $A \subseteq B \subseteq C$. □

[12] The non-monotonicity of *all but one* is obvious (provided that *one* is understood with an exact meaning, which is intuitively the only available reading in *all but one*; we thank an anonymous reviewer for pointing out this potential complication). Similarly, its non-symmetry is easily observed, e.g. 'All but one linguists are semanticists' is not equivalent to 'All but one semanticists are linguists'. It's left-continuity can be proved as follows, assuming that $[\![\text{all but one}]\!](A)(B)$ iff $|A| - |A \cap B| = 1$:

Proof. Suppose that $|A| = n$ for some $n > 0$, and $|A \cap X| = |C \cap X| = n - 1$. Then for any B, if $A \subseteq B$, then $|B \cap X| \geqslant n - 1$, and if $B \subseteq C$, then $|B \cap X| \leqslant n - 1$. Thus if $A \subseteq B \subseteq C$, then $|B \cap X| = n - 1$. □

(17c) together entail the ∀-reading of (17b).

Therefore, if it turns out that there is no preference between ∀-reading and ∃-reading with *all but one*, this will fit naturally with the first variant of Kanazawa's (1994) hypothesis; if it turns out that the preferred reading of *all but one* is the ∀-reading, this will fit naturally with the latter two variants; on the other hand, if it turns out that the ∃-reading is preferred, this will be at odds with all three variants. Table 1 summarizes the predictions of the three variants of Kanazawa's (1994) hypothesis.

	exactly three	*all but one*
1. symmetry + monotonicity	∃	No preference
2. left-continuity (+ monotonicity)	∃	∀
3. symmetry + left-continuity (+ monotonicity)	∃	∀

Table 1: Predictions of the three variants of Kanazawa's (1994) hypothesis for the two non-monotonic quantifiers, *exactly three* and *all but one*.

3 Experiment 1: *Exactly Three*

In Experiment 1, we explore the interpretation of donkey anaphora in the scope of the non-monotonic quantifier *exactly three*. What are the logically possible interpretations of donkey anaphora in such environments? To answer this question, we first note that the meaning of a non-monotonic quantifier can be decomposed into an upward monotonic and a downward monotonic component. For instance, *exactly three* is semantically equivalent to the conjunction of the upward entailing quantifier *at least three* with the downward entailing quantifier *at most three*. Given this, on the assumption that the donkey pronoun can get a ∃-reading or ∀-reading in each of the upward and downward meaning components, there are in principle four logically possible readings for a donkey pronoun in the scope of *exactly three*. That is, the pronoun could get (i) a ∃-reading in both components, (ii) a ∀-reading in both components, (iii) an ∃-reading in the upward component combined with a ∀-reading in the downward component and (iv) a ∀-reading in the upward component combined with a ∃-reading in the downward component. We will label these readings as (i) $\frac{\exists}{\exists}$, (ii) $\frac{\forall}{\forall}$, (iii) $\frac{\forall}{\exists}$, (iv) $\frac{\exists}{\forall}$, respectively. The mnemonic is that what's above the line is the reading of the upward component of the meaning and what's below the line is the reading of the downward component of the meaning. The four logically possible readings for (18) are paraphrased below.

(18) Exactly three squares that are above a heart are connected to it.

- $\frac{\exists}{\exists}$ At least three squares that are above a heart or hearts are connected to **some** of those hearts and at most three squares that are above a heart or hearts are connected to **some** of those hearts.
- $\frac{\forall}{\forall}$ At least three squares that are above a heart or hearts are connected to **all** of those hearts and at most three squares that are above a heart or hearts are connected to **all** of those hearts.
- $\frac{\forall}{\exists}$ At least three squares that are above a heart or hearts are connected to **all** of those hearts and at most three squares that are above a heart or hearts are connected to **some** of those hearts.

$\begin{smallmatrix}\exists\\ \forall\end{smallmatrix}$ At least three squares that are above a heart or hearts are connected to **some** of those hearts and at most three squares that are above a heart or hearts are connected to **all** of those hearts.

These four possible readings stand in the entailment relation depicted in Figure 1.

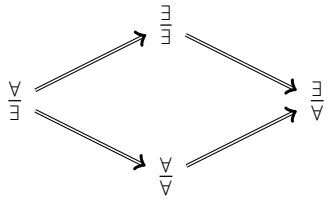


Figure 1: Entailments among the four logically possible readings

It is important to note that $\begin{smallmatrix}\exists\\ \exists\end{smallmatrix}$ corresponds to what is typically referred to as the $\exists$-reading of donkey anaphora; $\begin{smallmatrix}\forall\\ \forall\end{smallmatrix}$ corresponds to the $\forall$-reading; $\begin{smallmatrix}\forall\\ \exists\end{smallmatrix}$ corresponds to the 'conjunctive reading' discussed by Champollion et al.'s (2019) (i.e. the conjunction of the $\exists$ and $\forall$ readings). Decomposing non-monotonic quantifiers in an upward and a downward meaning component thus provides another perspective on relating the conjunctive reading of donkey pronoun to the $\exists$-reading and $\forall$-reading. Specifically, the conjunctive reading can be thought of as a *mixed* reading in the sense that it has different quantificational forces in the two meaning components of the non-monotonic quantifiers. On the other hand, $\begin{smallmatrix}\exists\\ \forall\end{smallmatrix}$ is also a mixed reading in this sense, but it does not correspond to a reading of donkey anaphora that has been previously discussed in the literature: we investigate it on a par with the other three readings that have been discussed previously in the literature for completeness.

To summarize this discussion, in Experiment 1, we investigate which of these four readings of donkey anaphora are available in the scope of the non-monotonic quantifier *exactly three*.

3.1 Task

Participants were directed to a web-based truth value judgment task, hosted on Alex Drummond's Ibex platform for psycholinguistic experiments. They were told that they would see sentences paired with images and that their task was to decide whether the sentence was true with respect to the image with which it was paired. They were instructed to record their responses on a bounded continuous scale, whose ends were labeled as 'Completely false' and 'Completely true'.

The participants first saw three practice trials, one involving a true sentence, one involving a false sentence, and one involving a sentence whose truth is harder to assess because it contained a vague quantifier *many*; these practice trials were accompanied by suggested responses. The purpose of these examples was to familiarize the participants with the task. They then began the test phase of the experiment, the first three items of which were identical to the three practice trials.

3.2 Materials

Sentences in Experiment 1 were always of the following form:

(19) Exactly three ⟨squares, triangles⟩ that are above a ⟨star, heart⟩ are connected to it.

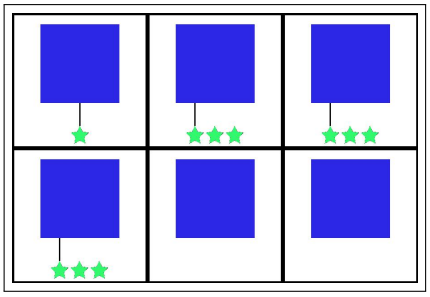
Participants' task was to judge whether such sentences are true with respect to an image; examples of images such sentences were matched with are in Figure 2.

Given the entailment relations among the four possible readings, there are four kinds of situations where at least one of these readings is true, which constitute our target conditions:

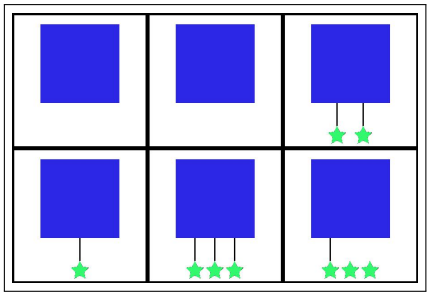
- DEweak-UEweak: Only the weakest reading $\frac{\exists}{\forall}$ is true, e.g. Figure 2a.
- DEweak-UEstrong: Only $\frac{\forall}{\forall}$ and $\frac{\exists}{\forall}$ are true, e.g. Figure 2b.
- DEstrong-UEweak: Only $\frac{\exists}{\exists}$ and $\frac{\exists}{\forall}$ are true, e.g. Figure 2c.
- DEstrong-UEstrong: All four readings are true, e.g. Figure 2d.

In addition, two control conditions were included in the experiment, where none of the four readings were true. We refer to them as DEfalse-UEstrong and DEstrong-UEfalse. DEfalse-UEstrong makes the upward entailing part of the meaning true under the ∀-reading, but falsifies the downward entailing part under both readings (e.g. Figure 2e). DEstrong-UEfalse makes the downward entailing part of the meaning true under the ∃-reading, but falsifies its upward entailing part under both readings (e.g. Figure 2f).

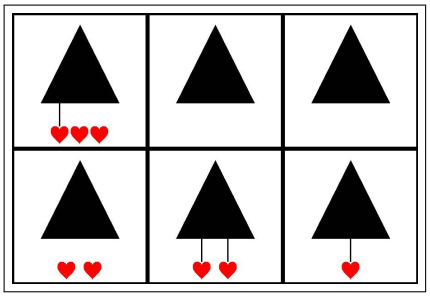
This amounts to a total of six conditions (four target and two control conditions). Since each of the six conditions had six items, there were 36 experimental items in total. For reasons that will be explained in Section 5, each participant first saw the items from DEstrong-UEstrong (six items), from DEstrong-UEweak (six items), and six items from false controls (3 items from DEstrong-UEfalse and three items from DEfalse-UEstrong). The order of these 18 items was randomized for each participant. These were followed by the remaining 18 items, the order of which was randomized for each participant as well.



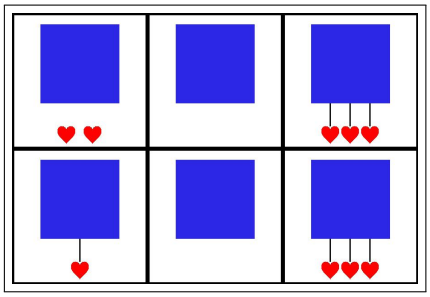
(a) DEweak-UEweak



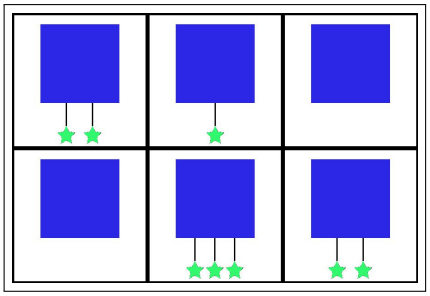
(b) DEweak-UEstrong



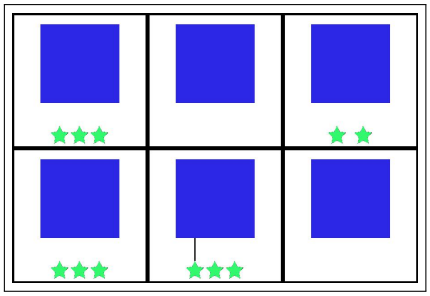
(c) DEstrong-UEweak



(d) DEstrong-UEstrong



(e) DEfalse-UEstrong



(f) DEstrong-UEfalse

Figure 2: Examples of experimental items images in six different conditions.

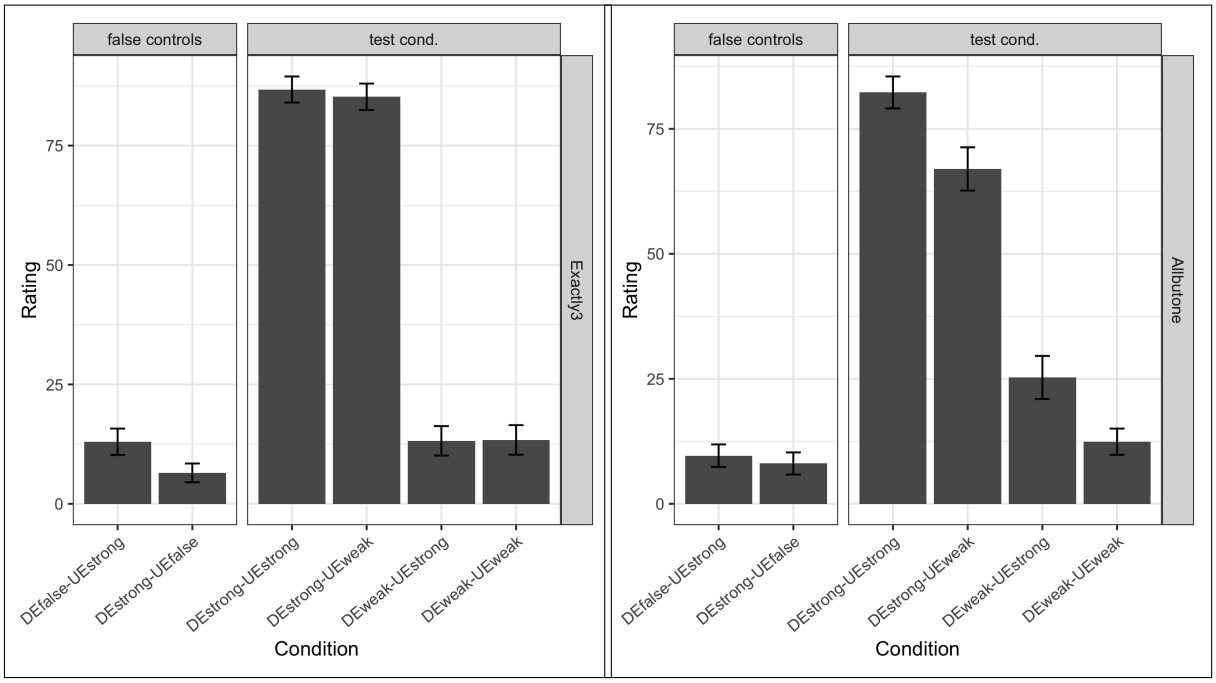
Each image consisted of six vignettes as in Figures 2. Each of the vignettes contained a large shape of the same kind (either triangles or squares). In four out of six vignettes the square/triangle was above one, two, or three instances of smaller shapes of the same kind (either stars or a hearts). There were thus two vignettes in which the square/triangle was not above any hearts/stars, which ensured the felicity of the relative clause in experimental sentences as in (19). In at least one of the vignettes the square or the triangle would appear above exactly one heart or star: this was to ensure the felicity of the singular morphology on the indefinite noun in the experimental sentences. For each item, a combination of shapes was chosen randomly (i.e. squares+stars, squares+hearts, triangles+stars, triangles+hearts), and the positions of the two vignettes with squares/triangles with no stars/hearts below them were chosen randomly as well. Likewise, the exact number of stars/hearts (one, two, or three) that appeared below the four squares/triangles in an item was chosen randomly for each of the four squares/triangles for each item, granting however that at least one of the squares/triangles would be above exactly one star/heart for felicity reasons mentioned above. We opted for having four squares/triangles that are above a star/heart in all of the experimental conditions because this permitted us to use the exact same visual stimuli for this experiment as for Experiment 2 that tested *all but one*.

3.3 Participants and exclusion criteria

65 participants (21 females) were recruited on Amazon Mechanical Turk. One participant was excluded for not being a native speaker of English. We furthermore excluded those participants whose average judgment in the four target conditions combined was not higher than their average judgment in the two control conditions combined. The logic behind this exclusion criterion is the following. If participants were able to access at least one of the four aforementioned logically possible readings, this should suffice for them to judge the target conditions on average better than the control conditions. If they did not do so, they might have not understood the experimental task, or they were possibly only able to access the uniqueness reading which was not verified in any of the six conditions and hence was not relevant for our purposes (cf. fn.2). This led to the exclusion of two additional participants. The remaining 62 participants were thus kept for the analyses.

3.4 Results

The results obtained are summarized in Figure 3a and Table 3. Recall that the target conditions render different logically possible readings true, as summarized in Table 2. Based on this, we will now discuss which readings the results give evidence for.



(a) Experiment 1: Exactly three (b) Experiment 2: All but one

Figure 3: Results of the two experiments per CONDITION. Error bars represent standard errors.

Reported p-values in Experiment 1 are adjusted for multiple comparisons by the Holm-Bonferroni method.

No evidence for $\overset{\exists}{\forall}$ If the weakest reading, $\overset{\exists}{\forall}$, has been accessed, DEWEAK-UEWEAK, which validates $\overset{\exists}{\forall}$, should receive higher rating than the control conditions, which validate none of the readings. The data was subsetted to the items in DEFALSE-UESTRONG and

	$\frac{\forall}{\exists}$	$\frac{\exists}{\exists}$	$\frac{\forall}{\forall}$	$\frac{\exists}{\forall}$
DEstrong-UEstrong	T	T	T	T
DEstrong-UEweak	F	T	F	T
DEweak-UEstrong	F	F	T	T
DEweak-UEweak	F	F	F	T
DEfalse-UEstrong	F	F	F	F
DEstrong-UEfalse	F	F	F	F

Table 2: Target conditions and the readings that they make true.

	Mean rating (SE)	
Condition	*Exp. 1: Exactly three*	*Exp. 2: All but one*
DEweak-UEweak	13.4 (3.1)	12.4 (2.6)
DEweak-UEstrong	13.2 (3.1)	25.3 (4.3)
DEstrong-UEweak	85.2 (2.8)	67 (4.3)
DEstrong-UEstrong	86.7 (2.7)	82.3 (3.2)
DEfalse-UEstrong	13 (2.8)	9.6 (2.3)
DEstrong-UEfalse	6.5 (2)	8.1 (2.2)

Table 3: Experiments 1 and 2: Mean participants' rating and standard error per condition.

DEweak-UEweak.[13] A linear mixed model was fitted on this data set with CONDITION as a fixed effect and random by-participant intercepts and slopes. A comparison of this model with a reduced model without CONDITION as a fixed effect revealed no significant effect of CONDITION ($\chi(1) = 0.06, p = 1$). There is thus no evidence for the existence of $\frac{\exists}{\forall}$ reading with *exactly three*.

No evidence for $\frac{\forall}{\forall}$ If $\frac{\forall}{\forall}$ has been accessed, DEweak-UEstrong, which validates both $\frac{\forall}{\forall}$ and $\frac{\exists}{\forall}$, should receive higher rating than DEweak-UEweak, which only validates $\frac{\exists}{\forall}$. The data was subsetted to the items in DEweak-UEstrong and DEweak-UEweak. A linear mixed model was fitted on this data set with CONDITION as a fixed effect and random by-participant intercepts and slopes. A comparison of this model with a reduced model without CONDITION as a fixed effect revealed no significant effect of CONDITION ($\chi(1) = 0.02, p = 1$). There is thus no evidence for the existence of $\frac{\forall}{\forall}$ reading with *exactly three*.

Evidence for $\frac{\exists}{\exists}$ If $\frac{\exists}{\exists}$ has been accessed, DEstrong-UEweak, which validates both $\frac{\exists}{\exists}$ and $\frac{\exists}{\forall}$, should receive higher rating than DEweak-UEweak, which only validates $\frac{\exists}{\forall}$. The data was subsetted to items in DEstrong-UEweak and DEweak-UEweak. A linear mixed model was fitted on this data set with CONDITION as a fixed effect and random by-participant intercepts and slopes. A comparison of this model with a reduced model without CONDITION as a fixed effect revealed a significant effect of CONDITION

[13] DEfalse-UEstrong was chosen rather than DEstrong-UEfalse because the mean rating of DEfalse-UEstrong was higher than that of DEstrong-UEfalse, and thus provides a stricter requirement for the detection of $\frac{\exists}{\forall}$.

($\chi(1) = 100, p < .001$). Our results thus provide evidence for the existence of $\frac{\exists}{\exists}$ reading with *exactly three*.

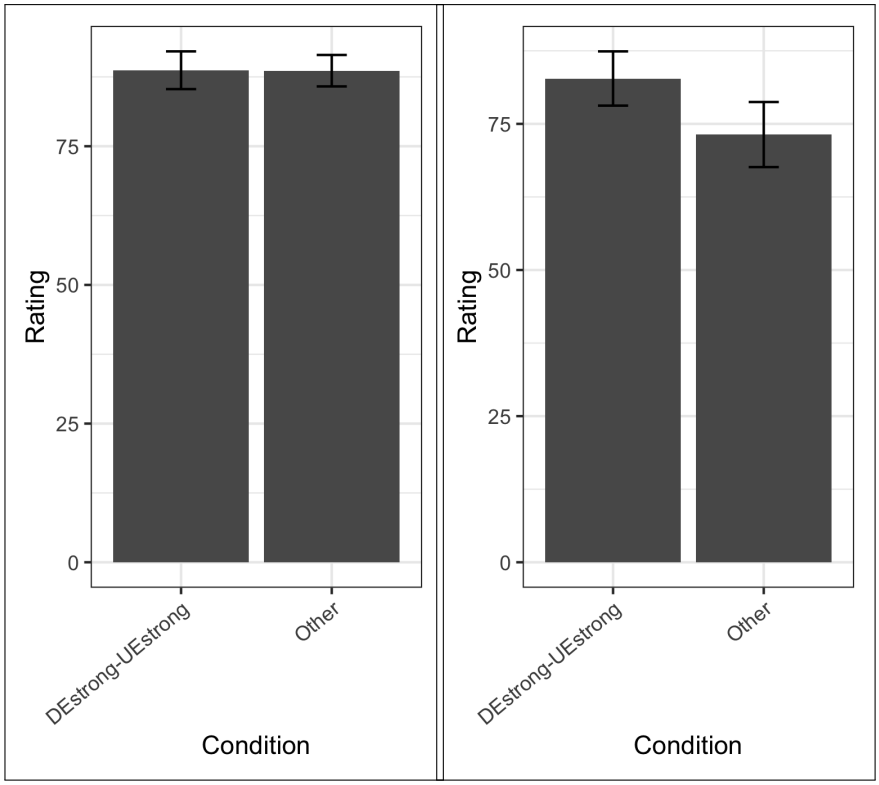
No evidence for $\frac{\forall}{\exists}$ In order to uncover whether the $\frac{\forall}{\exists}$ reading is available or not, we cannot simply compare the DEstrong-UEstrong condition, which is the only condition validating $\frac{\forall}{\exists}$, to some other condition. The reason is the following. Suppose the reading $\frac{\forall}{\exists}$ is never accessed, while the other three readings (i.e. $\frac{\forall}{\forall}$, $\frac{\exists}{\exists}$, and $\frac{\exists}{\forall}$) are accessed at least to some extent. This would mean that DEstrong-UEstrong validates all of the three available readings, while all the other conditions validate at most a proper subset thereof. This on its own might suffice to make participants rate items in DEstrong-UEstrong higher than in any of the remaining conditions. Therefore, a significant difference between DEstrong-UEstrong and any of the other conditions in itself would not constitute evidence for the existence of $\frac{\forall}{\exists}$ reading.

To circumvent this issue, we selected participants with the following property: their mean rating in at least one of DEstrong-UEweak and DEweak-UEstrong is equal or lower than in DEweak-UEweak. The idea is that these participants accessed at most one of $\frac{\exists}{\exists}$ and $\frac{\forall}{\forall}$. In other words, they (at most) accessed either (i) $\frac{\forall}{\forall}$, $\frac{\forall}{\exists}$, and $\frac{\exists}{\forall}$ or (ii) $\frac{\exists}{\exists}$, $\frac{\forall}{\exists}$, and $\frac{\exists}{\forall}$. This further means that, for the participants who did not access $\frac{\forall}{\forall}$, the only reading which is true in DEstrong-UEstrong but not in DEstrong-UEweak is $\frac{\forall}{\exists}$. Likewise, for participants who did not access $\frac{\exists}{\exists}$, the only reading which is true in DEstrong-UEstrong but not in DEweak-UEstrong is $\frac{\forall}{\exists}$. Thus, if these participants would rate DEstrong-UEstrong even better than the one they rated better between DEstrong-UEweak and DEweak-UEstrong, this could be taken as evidence that these participants accessed $\frac{\forall}{\exists}$. 42 participants fell into this category, and the following analysis was conducted on their responses.

The data was subsetted to items in DEstrong-UEstrong and the better rated condition between DEstrong-UEweak and DEweak-UEstrong (as determined for each participant separately). A linear mixed model was fitted on this data set with condition (DEstrong-UEstrong vs. Other) as a fixed effect and random by-participant intercepts and slopes. A comparison of this model with a reduced model without condition as a fixed effect revealed no significant effect of condition ($\chi(1) = 0.01, p = 1$). For reference, the mean rating of the better rated conditions between DEstrong-UEweak and DEweak-UEstrong (as determined for each participant separately) was 88.6 ($SE = 2.8$), while their mean rating in DEstrong-UEstrong was 88.7 ($SE = 3.4$) (cf. Figure 4a). There is thus no evidence for the existence of $\frac{\forall}{\exists}$ with *exactly three*.[14]

Summarizing the results of Experiment 1, the only detected reading of donkey anaphora in the scope of *exactly 3* is $\frac{\exists}{\exists}$ (i.e. $\exists$-reading).

[14] An anonymous reviewer asks whether it is possible to identify a sub-population whose ratings in both DEstrong-UEweak and DEweak-UEstrong conditions are low, suggesting that they access neither $\frac{\exists}{\exists}$ nor $\frac{\forall}{\forall}$ reading, but whose ratings in DEstrong-UEstrong condition are high, which may suggest that they access the $\frac{\forall}{\exists}$ reading. We have thus looked into whether there are any participants whose ratings in the DEweak-UEstrong condition and DEstrong-UEweak condition are equally low or lower than in the DEweak-UEweak condition, but their ratings in the DEstrong-UEstrong condition are higher than in the DEweak-UEweak condition. There is only one participant whose response pattern fits the above description, making it difficult to draw any strong conclusions about the existence of such a sub-population with the data collected in Experiment 1.



(a) Experiment 1: Exactly three (b) Experiment 2: All but one

Figure 4: Results of the two experiments for the participants who access at most one of $\frac{\exists}{\exists}$ and $\frac{\forall}{\forall}$ (possibly in addition to $\frac{\forall}{\exists}$ and/or $\frac{\exists}{\forall}$). The plot represents their mean rating of DEstrong-UEstrong and whichever was higher between their mean ratings of DEstrong-UEweak and DEweak-UEstrong (coded as Other). Error bars represent standard errors.

4 Experiment 2: *All but One*

4.1 Task and materials

Experiment 2 had the exact same task and materials as Experiment 1, except that the experimental sentences used *all but one*, in place of *exactly three* as in (20).

(20) All but one of the ⟨squares, triangles⟩ that are above a ⟨star, heart⟩ are connected to it.

As mentioned above, we constructed the pictures for Experiment 1 in such a way that they can be used in Experiment 2 as well. Thus, as the visual stimuli are kept constant in Experiments 1 and 2, any differences between the results of the two experiments cannot be a side-effect of the visual stimuli, but would have to be explained by the difference in linguistic stimuli, and/or the interaction between the linguistic and visual stimuli.

4.2 Participants and exclusion criteria

The procedure was identical to Experiment 1. A new set of 65 participants (25 females) were recruited on Amazon Mechanical Turk, none of whom participated in Experiment 1. One participant was excluded for failing to complete the experiment, two participants

were excluded for not being native speakers of English, and six participants were excluded for their average judgment in target conditions not being higher than their average judgment in control conditions (which is the same exclusion criterion as in Experiment 1). 56 participants were thus kept for the analysis.

4.3 Results

The results obtained are summarized in Figure 3b and Table 3. The logic of the data analysis is identical to that in Experiment 1, and we conducted parallel statistical analyses as follows.

Reported p-values in Experiment 2 are adjusted for multiple comparisons by the Holm-Bonferroni method.

No evidence for $\begin{smallmatrix}\exists\\ \forall\end{smallmatrix}$ Statistical analyses on data from DEweak-UEweak and DEfalse-UEstrong revealed no significant effect of condition ($\chi(1) = 2.35, p = .12$). There is thus no evidence for the existence of $\begin{smallmatrix}\exists\\ \forall\end{smallmatrix}$ with *all but one*.

Evidence for $\begin{smallmatrix}\forall\\ \forall\end{smallmatrix}$ Statistical analyses on data from DEweak-UEweak and DEweak-UEstrong showed that unlike in Experiment 1, DEweak-UEstrong was judged significantly better than DEweak-UEweak ($\chi(1) = 10.5, p < .01$). This result provides evidence for the existence of $\begin{smallmatrix}\forall\\ \forall\end{smallmatrix}$ with *all but one*.

Evidence for $\begin{smallmatrix}\exists\\ \exists\end{smallmatrix}$ Statistical analyses on data from DEweak-UEweak and DEstrong-UEweak indicate that as in Experiment 1, DEstrong-UEweak is judged significantly better than DEweak-UEweak ($\chi(1) = 64.8, p < .001$). This result provides evidence for the existence of $\begin{smallmatrix}\exists\\ \exists\end{smallmatrix}$ with *all but one*.

No evidence for $\begin{smallmatrix}\forall\\ \exists\end{smallmatrix}$ As in Experiment 1, in order to determine whether participants have accessed $\begin{smallmatrix}\forall\\ \exists\end{smallmatrix}$, we selected participants whose mean rating in at least one of DEstrong-UEweak and DEweak-UEstrong is equal or lower than in DEweak-UEweak. 33 participants fell into this category in Experiment 2. Analyses parallel to Experiment 1 were conducted on their responses in DEstrong-UEstrong and the better rated condition between DEstrong-UEweak and DEweak-UEstrong (as determined for each participant separately). They revealed no significant effect of condition (DEstrong-UEstrong vs. other) ($\chi(1) = 3.62, p = .11$). For reference, the mean rating of the better rated conditions between DEstrong-UEweak and DEweak-UEstrong (as determined for each participant separately) was 73.2 ($SE = 5.6$), while the mean rating in DEstrong-UEstrong was 82.3 ($SE = 4.6$) (cf. Figure 4b).[15]

Summarizing the results of Experiment 2, we detected $\begin{smallmatrix}\exists\\ \exists\end{smallmatrix}$ (i.e. the $\exists$-reading) and $\begin{smallmatrix}\forall\\ \forall\end{smallmatrix}$ (i.e. the $\forall$-reading) of donkey anaphora in the scope of *all but one*. Between these two, the $\exists$-reading is nonetheless preferred to the $\forall$-reading, as the DEstrong-UEweak condition

[15] As in Experiment 1, we have looked into whether there are any participants whose ratings in the DEweak-UEstrong condition and DEstrong-UEweak condition are equally low or lower than in the DEweak-UEweak condition, but their ratings in the DEstrong-UEstrong condition are higher than in the DEweak-UEweak condition. There is again only one participant whose response pattern fits the above description, making it difficult to draw any strong conclusions with the data collected in Experiment 2 about whether there is a sub-population which accesses neither $\begin{smallmatrix}\exists\\ \exists\end{smallmatrix}$ nor $\begin{smallmatrix}\forall\\ \forall\end{smallmatrix}$ reading, but may access the $\begin{smallmatrix}\forall\\ \exists\end{smallmatrix}$.

is judged significantly better than the DEweak-UEstrong condition, as evidenced by analyses parallel to those reported above ($\chi(1) = 26.8, p < .001$).

5 Experiment 3A: *All*

Before we discuss the findings of Experiments 1 and 2 further, we describe the results of an additional experiment investigating donkey anaphora interpretation with the universal quantifier *all*. The motivation for conducting Experiment 3 was two-fold. Firstly, Experiment 3 addresses a potential reproach to the generalizability of findings of Experiment 1 and 2. Namely, the fact that only $\frac{\exists}{\exists}$ was detected with *exactly three* and that $\frac{\exists}{\exists}$ was clearly the preferred one with *all but one* as well raises the question of whether something about our experimental setup might be biasing strongly towards $\frac{\exists}{\exists}$ (i.e. the $\exists$-reading), and thus masking the existence of, or even preference for, other readings, such as for instance the $\forall$-reading ($\frac{\forall}{\forall}$) with *exactly three* and $\frac{\forall}{\exists}$ with both quantifiers. The results of Experiment 3 attest that this is not the case. In the truth value judgment task of Experiment 3, which we will refer to as Experiment 3A, whose setting is comparable to Experiments 1 and 2, we find that the rate of $\exists$-readings with *all* is comparable to the rates of this reading reported in other experimental studies [Foppolo, 2008, Sun et al., 2020], and it is significantly lower than the rate of $\exists$-readings (i.e. $\frac{\exists}{\exists}$) for *all but one* and *exactly 3* (see Section 5.6 for relevant analyses). Even though we cannot claim that our experimental items are introducing absolutely no bias towards $\exists$-reading whatsoever, the considerations above suggest there is no reason to believe that in Experiments 1 and 2 we are greatly overestimating it, or that this bias is the sole culprit for absence or low rates of readings other than $\exists$-reading ($\frac{\exists}{\exists}$) with *all but one* or *exactly 3*.

The second motivation for Experiment 3 was to investigate whether the preservation of subjective inferential patterns matters for the interpretation of donkey anaphora. To this end, in addition to the truth value judgment task, Experiment 3 also included an inference judgment task, which we will refer to as Experiment 3B. Investigating simultaneously the results of the truth value judgment task and of the inference judgment task will allow us to evaluate whether participants' perceived monotonicity and symmetry properties of the quantifier *all* explain the extent to which the $\forall$-reading and $\exists$-reading of donkey anaphora are available with this quantifier. This second investigation effectively led to null results; we have thus opted to describe Experiment 3B in the Appendix.

5.1 Task

Experiment 3 had two tasks: a truth value judgment task, which was administered first, followed by an inference judgment task. We describe the former here (Experiment 3A); the latter is described in Appendix (Experiment 3B). The truth value judgment task was administered first in order for the participants of Experiment 3 to complete it in similar circumstances as the participants of Experiments 1 and 2.

The instructions and practice items for the truth value judgment task of Experiment 3 were identical to those in Experiments 1 and 2.

5.2 Materials

Truth value judgment task Sentences in this task were always of the form as in (21):

(21) All of the ⟨squares, triangles⟩ that are above a ⟨star, heart⟩ are connected to it.

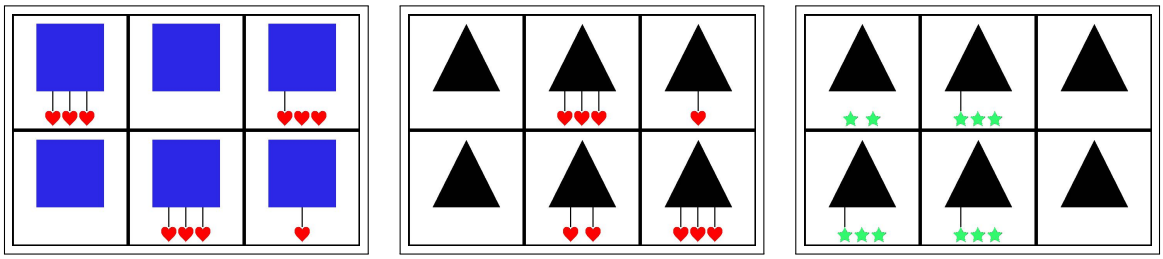
Just like in Experiments 1 and 2, participants' task was to judge whether such sentences are true with respect to an image; examples of images such sentences were matched with are in Figure 5.

There were two target conditions, corresponding to two logically possible types of situations in which at least one of the ∀-reading and ∃-reading is true (note that the ∀-reading entails the ∃-reading):

- Weak, in which only the ∃-reading is true, cf. Figure 5a.
- Strong, in which both ∀-reading and ∃-reading are true, cf. Figure 5b.

In addition, there was one control condition (False), in which neither of the two readings were true (cf. Figure 5c). This amounts to a total of three conditions (two target and one control). Since each of the three conditions had six items, there were thus 18 items in total in the truth value judgment task of Experiment 3. These 18 items were presented in a randomized order for each participant.

Images employed in this task were subject to similar constraints as in Experiments 1 and 2 (cf. Section 3.2).



(a) Weak (b) Strong (c) False

Figure 5: Examples of experimental items in Experiment 3A.

5.3 Participants and exclusion criteria

65 participants (15 females) were recruited on Amazon Mechanical Turk, none of whom participated in Experiment 1 or Experiment 2. All participants reported being native speakers of English. One participant was excluded for failing to complete the experiment, and additional four participants were excluded because (i) their average judgment in the target conditions was not higher than their average judgment in the control conditions in the truth value judgment task (which is the same exclusion criterion as in Experiments 1 and 2), or (ii) their average judgment on the valid control conditions was not higher than their average judgment on the invalid control conditions in the inference judgment task (cf. Appendix for details about the controls in the inference judgment task). 60 participants were thus kept for the analysis.

5.4 Results summary

The results obtained in the truth value judgment task are summarized in Figure 6 and Table 4.

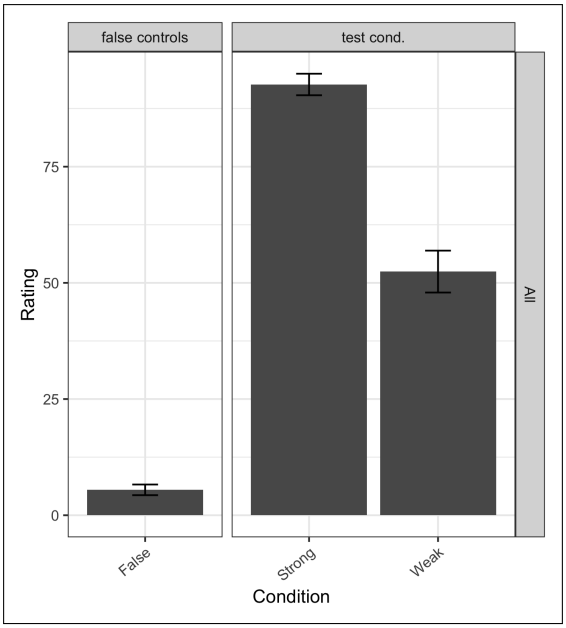


Figure 6: Results of the truth value judgment task of Experiment 3 per CONDITION (Experiment 3A). Error bars represent standard errors.

Condition	Mean rating (SE)
STRONG	92.7 (2.3)
WEAK	52.4 (4.5)
FALSE	5.5 (1.1)

Table 4: Experiment 3A: Mean participants' rating and standard error per condition in the truth value judgment task.

5.5 ∃-reading and ∀-reading of donkey anaphora

There is clear evidence for both ∃-reading and ∀-reading of donkey anaphora in the case of quantifier *all*, as evidenced by statistical analyses parallel to those for Experiments 1 and 2.

Reported p-values in Experiment 3A are adjusted for multiple comparisons by the Holm-Bonferroni method.

Evidence for ∃-reading Statistical analyses on data from the FALSE and WEAK conditions revealed a significant effect of CONDITION ($\chi(1) = 67.7, p < .001$). There is thus evidence for the ∃-reading with *all*.

Evidence for ∀-reading Statistical analyses on data from the STRONG and WEAK conditions revealed a significant effect of CONDITION ($\chi(1) = 49.2, p < .001$). There is thus evidence for the ∀-reading with *all*.

5.6 Comparison with Experiments 1 and 2

The design of experiment 3A was parallel to that of Experiments 1 and 2, and this allows us to discern whether the specifics of our experimental setup are to blame for

the prevalence of ∃-reading with *all but one* and *exactly 3*. Recall that the first 18 items of Experiments 1 and 2 consisted of six items from DEstrong-UEstrong, six items from DEstrong-UEweak, and six items from the false controls. They were thus comparable to the 18 items of Experiment 3A, of which six validated both ∀- and ∃-reading of donkey anaphora, six validated only the ∃-reading, and six didn't validate either. (The comparability stems from the fact that, modulo $\frac{\forall}{\exists}$ and $\frac{\exists}{\forall}$, DEstrong-UEstrong validates both ∀- and ∃-reading of donkey anaphora, DEstrong-UEweak validates only the ∃-reading, and the false controls do not validate either.) If the specifics of the experimental setup are biasing strongly towards the ∃-reading in Experiments 1 and 2, we expect to observe this bias in Experiment 3A as well. As is evident from the results of Experiment 3A with quantifier *all* (cf. Figure 6), the condition which makes the ∃-reading but not the ∀-reading true is judged significantly lower in Experiment 3A than in Experiments 1 and 2. This is confirmed by a significant Quantifier-Condition interaction: *All vs. All but one*: $(\chi(1) = 14.7, p < .01)$; *All vs. Exactly 3*: $(\chi(1) = 53.6, p < .01$.[16] In other words, ∃-reading is available significantly less with *all* than with *all but one* or *exactly 3*.

It is also relevant to point out that other experimental work with picture verification-based truth value judgment tasks which investigated ∃- and ∀-readings with donkey anaphora with quantifier *all* mostly found comparable rates of ∃-readings as we did in Experiment 3A. In Experiment 1 of Foppolo [2008], the condition which validates both ∃- and ∀-readings with a universal quantifier is judged true 100% of the time, while the condition which validates the ∃-reading only is judged true 57% of the time. In Experiment 1 reported in Sun et al. [2020], the condition which validates both ∃- and ∀-readings with a universal quantifier is judged true approximately 90% of the time, while the condition which validates the ∃-reading only is judged true approximately 60% of the time. A potential outlier is the results reported in Geurts [2002], in which the condition which validates the ∃-reading only are judged true on average 38% of the time, with a large variation between different sub-conditions. Geurts [2002] however did not test the condition in which both ∃- and ∀-readings are true, which leaves open the possibility that his subjects were rejecting the test sentences in the condition which validates the ∃-reading for reasons other than the access to the ∀-reading, and makes it difficult to draw firm conclusions from his results about the the rate of ∃-readings with a universal quantifier.

These two facts (that the rate of ∃-readings with *all* we observe is comparable to the rate observed in other studies; and that the ∃-reading is significantly less robustly available with *all* than with *all but one* or *exactly 3*) suggest that the low rates of readings other than the ∃-reading with *all but one* and *exactly three* are due to those specific quantifiers rather than due to the experimental setup biasing strongly towards ∃-reading. It is however not possible to exclude that some bias may nonetheless be present in our experimental setup; the reading rates observed in our experiments should thus be interpreted as a rough approximation of participants' natural preferences.

[16] In both cases this was established as follows: a linear mixed model was fitted on responses from Strong and Weak for quantifier *all*, and responses from DEstrong-UEstrong and DEstrong-UEweak which were re-coded as Strong and Weak respectively, with quantifier, condition, and the quantifier-condition interaction term as fixed effects, and random by-participant intercepts and slopes. This model was then compared with a reduced model without the quantifier-condition interaction term.

6 General discussion

Let us summarize the main empirical findings from Experiments 1 and 2.

1. We obtained no evidence that $\frac{\exists}{\forall}$ is available with either *exactly three* or *all but one*. This is unsurprising; this reading, even though logically possible, has not been reported in the literature so far.

2. There is clear evidence that $\frac{\exists}{\exists}$ is available with both *exactly three* and *all but one*. According to our results, this reading is the most prominent one with both quantifiers.

3. Interestingly, we find differences between *all but one* and *exactly three* with respect to the availability of $\frac{\forall}{\forall}$.

4. Finally, we do not find conclusive evidence for the existence of $\frac{\forall}{\exists}$ with either *exactly 3* or *all but one*.

In the remainder of this section, we will discuss the theoretical implications of our results.

6.1 Symmetry vs. Left-Continuity

Recall the predictions of the three variants of Kanazawa's (1994) hypothesis (see Table 1).

- All three predict the default reading of *exactly three* is the $\exists$-reading, which we have been calling $\frac{\exists}{\exists}$.

- If left-continuity needs to be preserved, the default reading of *all but one* should be the $\forall$-reading, which we called $\frac{\forall}{\forall}$.

- If symmetry but not left continuity needs to be preserved, there shouldn't be a preference between the $\exists$-reading ($\frac{\exists}{\exists}$) and $\forall$-reading ($\frac{\forall}{\forall}$) with *all but one*.

In Experiment 1, we found no evidence for the $\forall$-reading with *exactly three*: the only observed reading with *exactly three* was the $\exists$-reading. This is consistent with the first point above, i.e. this finding is compatible with all three variants of Kanazawa's (1994) hypothesis. What about the results of Experiment 2? In Experiment 2, we found evidence for the availability of both the $\exists$-reading and $\forall$-reading with *all but one*, with the $\exists$-reading clearly preferred to the $\forall$-reading. That the $\exists$-reading is preferred for *all but one* is at odds with all three variants of Kanazawa's (1994) hypothesis, which predict either a preference for the $\forall$-reading (the two variants involving left-continuity) or no preference (the variant involving symmetry but not left-continuity), as summarized in Table 1.

Could the three variants of Kanazawa's (1994) hypothesis be amended to resolve the tension with our experimental results? If the symmetry variant of Kanazawa's (1994) hypothesis was on the right track (Variant 1 in Table 1), the tension with our experimental results could be resolved on the assumption that our experiments introduced a *minor* bias for the $\exists$-reading. This would suffice to explain the observed preference for $\exists$-reading with *all but one* in Experiment 2, and does not contradict the results of Experiment 3A.

On the other hand, if one of the left-continuity variants of Kanazawa's (1994) hypothesis were on the right track (Variants 2 and 3 in Table 1), resolving the tension with our

experimental results would force us to assume that our experiments introduced a *major* bias for the ∃-reading, which would entirely flip the natural reading preferences for *all but one*, from the predicted preference for ∀-reading to the observed preference for ∃-reading.

Recall that in Section 5 we provided arguments, based on experimental results about the rates of ∃-reading with the quantifier *all*, against the possibility that our experimental setting introduced a major bias for the ∃-reading, and pointed out that we cannot show that it had no bias whatsoever. Thus, our results are easier to make sense of under the symmetry variant of Kanazawa's (1994) hypothesis than within one of the left-continuity variants. That is, the version of Kanazawa's (1994) hypothesis that requires monotonicity and symmetry to be preserved fares better than the other two that require left-continuity to be preserved, even though it needs additional assumptions to predict that, while both ∃-reading and ∀-reading are clearly accessible with *all but one*, ∃-reading is nonetheless preferred in our experiments.

As an anonymous reviewer rightly pointed out, however, it should be noticed that our results do not constitute positive evidence for the symmetry+monotonicity hypothesis: it is superior to its contender hypotheses only because it is easier to accommodate with our experimental results than the latter. In order to gain direct evidence for the relevance of symmetry, it is necessary to investigate readings of donkey anaphora with non-left-continuous but symmetric non-monotonic quantifiers, such as *an odd number of, an even number of*, or *exactly three or exactly five.* We will leave this for future research.

6.2 The Conjunctive Reading and Pragmatic Factors

We now turn to a recent theory proposed by Champollion et al. [2019], mentioned briefly in Section 1 . According to them, donkey anaphora sentences are semantically true under the 'conjunctive reading', i.e. if both the ∃- and ∀-readings are true, false if both of them are false, and neither true nor false otherwise. Applying this recipe to *every* and *no*, for example, we get the following meaning:[17]

(22) 'Every farmer who owns a donkey beats it' is:

a. *true*, if every donkey-owning farmer beats all of their donkeys;
b. *false*, if at least one donkey-owning farmer beats none of their donkeys; and
c. *neither true nor false*, otherwise.

(23) 'No farmer who owns a donkey beats it' is:

a. *true*, if no donkey-owning farmer beats any of their donkeys;
b. *false*, if some donkey-owning farmer beats all of their donkeys; and
c. *neither true nor false*, otherwise.

In addition to this semantics, Champollion et al. [2019] assume that depending on various contextual factors (such as Questions under Discussion (QUD), cf. Roberts, 2012), a given donkey sentence might be perceived as true or false, even in cases where it is semantically neither true nor false. That is, contextual considerations may allow one to judge cases that are neither true nor false as simply true or simply false. Note in particular that for the above two examples, when the false and neither true nor false cases are judged as false, the resulting response pattern is indistinguishable from the

[17] Champollion et al. [2019] assume that indefinites like *a NP* and *some NP* are not quantifiers, and therefore they simply do not give rise to such trivalent meaning, and receive ∃-readings.

∀-reading in (22) and the ∃-reading in (23) in a bivalent setting. Alternatively, when the true and neither true nor false cases are judged as true, the resulting response pattern is indistinguishable from the ∃-reading in (22) and the ∀-reading in (23).

This theory would assign the following truth-conditions to the two non-monotonic quantifiers we tested.

(24) 'Exactly three farmers that own a donkey beat it.' is:

a. *true*, if there are three donkey-owning farmers who beat all of their donkeys and no other donkey-owning farmer beats any of their donkeys;

b. *false*, if the number of donkey-owning farmers who beat all of their donkeys is not three, and the number of donkey-owning farmers who beat at least one of their donkeys is not three; and

c. *neither true nor false*, otherwise.

(25) 'All but one farmer who owns a donkey beats it' is:

a. *true*, if there is one donkey-owning farmer who beats none of their donkeys and all the other donkey-owning farmers beat all of their donkeys;

b. *false*, if either there is more than one donkey-owning farmer who beats none of their donkeys, or all donkey-owning farmers beat all of their donkeys; and

c. *neither true nor false*, otherwise.

To put it differently, for both of these cases, the sentence is true when $\frac{\forall}{\exists}$ is true, and it is false when neither $\frac{\exists}{\exists}$ nor $\frac{\forall}{\forall}$ is true, and in cases where $\frac{\exists}{\exists}$ or $\frac{\forall}{\forall}$ is true but not $\frac{\forall}{\exists}$, the sentence is semantically neither true nor false. If the *neither true nor false* cases are judged as false, the resulting response pattern is indistinguishable from $\frac{\forall}{\exists}$ reading (i.e. the conjunctive reading).

The results of our experiment pose several challenges for this theory. Firstly, as we have discussed, we find no empirical evidence for the availability of the conjunctive reading, $\frac{\forall}{\exists}$, for sentences with donkey anaphora with *exactly three* or *all but one*. Of course, this does not in itself disprove Champollion et al.'s (2019) analysis, but to the extent that their theory predicts the availability of this reading, it would be ideal to raise empirical evidence for it.

Secondly, we obtained evidence for ∀-readings for *all but one* but not for *exactly three*. For *exactly three*, the only reading detected in our experiment was the ∃-reading. According to Champollion et al.'s (2019) theory, this means that the two quantifiers differ with respect to when neither true nor false sentences are judged as true. How can this difference between the two quantifiers be accommodated within Champollion et al. theory?

Champollion et al. [2019, § 6.3] discuss the possibility that symmetry/intersectivity is a relevant factor in judging a sentence that is neither true nor false as true. Specifically, they mention, although do not seem to necessarily countenance, the idea due to Geurts [2002] that symmetric/intersective quantifiers allow one to zoom in on the 'positive evidence' in the visual scene, and with respect to that part of the visual scene, the ∀- and ∃-readings simply collapse, making the conjunctive reading true, even if it is false with respect to the entire scene. For example, for (26), the relevant subpart of the visual scene contains all the triangles but not all the hearts: it only contains those hearts that are connected to a triangle.

(26) Exactly three triangles that are above a heart are connected to it.

As a consequence, the sentence containing a symmetric quantifier is accepted often in situations where what is traditionally called the ∃-reading is true, even when the conjunctive reading is false. Furthermore, by assumption, this strategy does not apply to non-intersective quantifiers. Given that *exactly three* is symmetric, while *all but one* is not (cf. Section 2), the response pattern indistinguishable from the ∃-reading may be more frequent with the former than with the latter. This creates room to capture the difference between the two quantifiers. However, to our knowledge, a precisely formulated and testable theory based on 'positive evidence' is yet to be worked out, which makes it difficult to evaluate with respect to our results.

Another potential way to make Champollion et al.'s (2019) theory compatible with the difference between *exactly three* and *all but one* in terms of the availability of the ∀-reading would be to assume that the two non-monotonic quantifiers preferred different types of QUDs in such a way that *exactly three* is consistently associated with QUDs that give rise to $\frac{\exists}{\exists}$, while *all but one* is also compatible with QUDs that give rise to $\frac{\forall}{\forall}$. In our experiments, the QUD is left implicit. To our knowledge, how participants infer the QUD when it is left implicit is still largely an open question. It is however indeed plausible that *exactly 3* and *all but one* sentences give rise to different inferred QUDs. Namely, there are constraints on the coherence between questions and answers in a dialogue which depend on the focus alternatives of answers [Rooth, 1992, von Stechow, 1991, Roberts, 2012]. If sentences headed by *exactly 3* activate different focus alternatives from sentences headed by *all but one*, it is expected that they would be compatible with different QUDs. We leave working out a theory that would connect these different QUDs to donkey anaphora interpretation for future work.

7 Conclusions and Further Directions

To summarize, we reported on two experiments that investigated which readings of donkey anaphora are available in the scope of two non-monotonic quantifiers, *exactly three* and *all but one*. To our knowledge, this is the first experimental study to investigate donkey anaphora interpretation in non-monotonic environments (but see Sun et al. 2020). Our results indicate that $\frac{\exists}{\exists}$ is available with both quantifiers, while we only obtained evidence of $\frac{\forall}{\forall}$ for the latter. Furthermore, $\frac{\exists}{\exists}$ was preferred to $\frac{\forall}{\forall}$ with *all but one*. We argued that these results speak against left-continuity playing a role in donkey anaphora interpretation. We also suggest that our results may be accommodated within a version of Kanazawa's (1994)'s hypothesis that symmetry needs to be preserved, together with monotonicity (but not left-continuity), under complementary assumptions which would predict that $\frac{\exists}{\exists}$ was preferred to $\frac{\forall}{\forall}$ with *all but one* in our experiments.

Can one provide stronger empirical support that either monotonicity or symmetry matters for the interpretation of donkey anaphora? In the aforementioned Experiment 3B, which we report in the Appendix, we ask whether participants' subjective perception of the monotonicity profile and of the symmetry profile of the quantifier *all* predicts the interpretation of donkey anaphora with that quantifier. We find no evidence that this is the case, however. This suggests that the main factor in interpretation of donkey anaphora is not the preservation of inferential patterns a participant would make with a given quantifier in sentences without donkey anaphora, as would be expected given Kanazawa's (1994) theory. In other words, if the logical properties of quantifiers are the main factor for donkey anaphora interpretation, this appears to be for reasons other than

the preservation of (subjective) inferential patterns with a given quantifier.

An alternative theory, according to which the monotonicity profile of the quantifier matters indirectly, is that of Champollion et al.'s (2019). Our results pose challenges for this theory as well by calling for evidence for the existence of $\frac{\forall}{\exists}$, and for an addition to their theory which would predict the observed differences between *all but one* and *exactly three* with respect to the availability of the ∀-reading.

A Experiment 3B: Subjective Monotonicity and Symmetry

Kanazawa's proposal according to which the preservation of inferential properties of quantifiers mediate the interpretation of donkey anaphora makes no commitments about how these inferential properties and donkey anaphora are jointly cognitively processed. Experiment 3B investigates a specific version of Kanazawa's proposal according to which some individual's interpretation of donkey anaphora hinges on that individual's subjective computation and representation of the inferential properties of the quantifiers.

This experiment is directly inspired by the experiment reported in Chemla, Homer, and Rothschild [2011], in which it was found that subjective inferential patterns matter in the case of negative polarity items licensing. In Experiment 3B, we focus on quantifier *all* and ask whether participants' perceived monotonicity and symmetry properties of this quantifier explain the extent to which the ∀-reading and ∃-reading of donkey anaphora are available with this quantifier.

To preview the findings, we find no correlation with either of the subjective versions of the two logical properties. This suggests that preservation of subjective inferential patterns that are due to either monotonicity or symmetry does not mediate reading preferences of donkey anaphora.

A.1 Tasks

As pointed out in Section 5, Experiment 3 had two tasks: a truth value judgment task, which was administered first, followed by an inference judgment task. The truth value judgment task of Experiment 3 is described in Section 5.

The procedure for the inference judgment task was as follows. Participants were told that they would see pairs of sentences about animals from the planet Zoopiter, and that Zoopiter is a planet similar to Earth, except for the fact that animals from Zoopiter have human-like hobbies and interests. Such a setting was chosen in order to minimize the influence of general world knowledge on inferences participants may derive. They were asked to evaluate to what extent the first sentence of the pair suggests that the second is true. Participants were instructed to record their responses on a bounded continuous scale, whose ends were labeled 'Not at all' and 'Very strongly'.

In the inference judgment task (as it was the case in the truth value judgment task), participants first saw three practice trials, one involving a case of a clearly valid inference, one involving a case of a clearly invalid inference, and one involving a case whose validity is harder to assess, accompanied by suggested responses. The purpose of these examples was to familiarize the participants with the task. They then began the test phase of the experiment, the first three items of which were identical to the three practice trials.

A.2 Materials

Inference judgment task At each experimental item, participants were presented with two sentences — a premise and a conclusion — and asked to evaluate to what extent the premise suggests that the conclusion is true. The premise was always a universally quantified sentence of the form *All X Y*.

This task had five target conditions: Symmetry, Restrictor-DE, Restrictor-UE, Scope-DE, Scope-UE. In addition, there were four control conditions which tested inferences of universally quantified sentences that were not of theoretical interest in the present study: two conditions included inferences that are valid (Valid Controls 1 and Valid Controls 2), and two inferences that are invalid (Invalid Controls 1 and Invalid Controls 2). Each condition had six items; there were thus 54 items in total in the inference judgment task of Experiment 3. These 54 items were presented in a randomized order for each participant.

For the items in Symmetry, participants had to evaluate to what extent (27a) suggests (27b) is true, with $[X, Y]$ taken from the list in (27c). Each of the pairs from (27c) appeared in exactly one of the six items in this condition.

(27) a. All of the Xs from Zoopiter are Ys.
b. All of the Ys are Xs from Zoopiter.
c. [[mosquito, cinema enthusiast], [dove, literature enthusiast], [poodle, music enthusiast], [crocodile, theater enthusiast], [tarantula, travel enthusiast], [cobra, museum enthusiast]]

For the items in Restrictor-DE, participants had to evaluate to what extent (28a) suggests (28b) is true; for items in Restrictor-UE, they had to evaluate to what extent (28b) suggests (28a). For these two conditions, $[X, Y]$ were taken from the list in (28c), and Z was taken from (28d). Each of the pairs from (28c) and each of the elements from (28d) appeared in exactly one of the six items in each of these two conditions.

(28) a. All of the Xs from Zoopiter Z.
b. All of the Ys from Zoopiter Z.
c. [spider, tarantula], [snake, cobra], [dog, poodle], [insect, mosquito], [bird, dove], [reptile, crocodile]
d. planted a rose, watched a documentary, wrote a novel, talked to a sculptor, travelled to France, prepared a Japanese dish

For the items in Scope-DE, participants had to evaluate to what extent (29a) suggests (29b) is true; for items in Scope-UE, they had to evaluate to what extent (29b) suggests (29a) is true. For these two conditions, $[Y, Z]$ were taken from the list in (29c), and X was taken from (29d). Each of the pairs from (29c) and each of the elements from (29d) appeared in exactly one of the six items in each of these two conditions.

(29) a. All of the Xs from Zoopiter Y.
b. All of the Xs from Zoopiter Z.
c. [planted a flower, planted a rose], [watched a film, watched a documentary], [wrote a book, wrote a novel], [talked to an artist, talked to a sculptor], [travelled to Europe, travelled to France], [prepared an Asian dish, prepared a Japanese dish]
d. poodle, mosquito, dove, crocodile, tarantula, cobra

For the items in Valid Controls 1, participants had to evaluate to what extent sentences such as (30a) suggest sentences such as (30b) are true. For the items in Valid Controls 2, participants had to evaluate to what extent sentences such as (30a) suggest sentences such as (30c) are true. For the items in Invalid Controls 1, participants had to evaluate to what extent sentences such as (30a) suggest sentences such as (30d) are true. For the items in Invalid Controls 2, participants had to evaluate to what extent sentences such as (30a) suggest sentences such as (30e) are true. Nouns appearing in the restrictor of the quantifier were taken from (29d), and predicates in the scope of the quantifier from (28d); each noun from (29d) and each predicate from (28d) appeared in exactly one of the six items in each of the four control conditions.

(30) a. All of the doves from Zoopiter planted a rose.
b. There is no dove from Zoopiter who didn't plant a rose.
c. There is a dove from Zoopiter who planted a rose.
d. Two doves from Zoopiter planted a rose.
e. Three doves from Zoopiter planted a rose.

A.3 Participants and exclusion criteria

These are described in Section 5.

A.4 Results summary

The results obtained in the inference judgment task are summarized in Figure 7 and Table 5.

Condition	Mean rating (SE)
Invalid controls 1	5.1 (1.6)
Invalid controls 2	5.6 (1.5)
Restrictor-DE	92.9 (1.5)
Restrictor-UE	27.7 (3.6)
Scope-DE	33.8 (3.6)
Scope-UE	91.7 (1.5)
Symmetry	29.1 (4)
Valid controls 1	91.7 (2.7)
Valid controls 2	89.8 (2.1)

Table 5: Experiment 3B: Mean participants' rating and standard error per condition in the inference judgment task.

A.5 Subjective monotonicity and the amount of ∀-reading

We first note that there is a noticeable variation in terms of how the participants judged the Weak condition in the truth value judgment task (cf. Experiment 3A in Section 5), as evidenced by the distribution of the participants' means in Figure 8. This shows that speakers differ in the extent to which they access the ∃-reading and the ∀-reading.

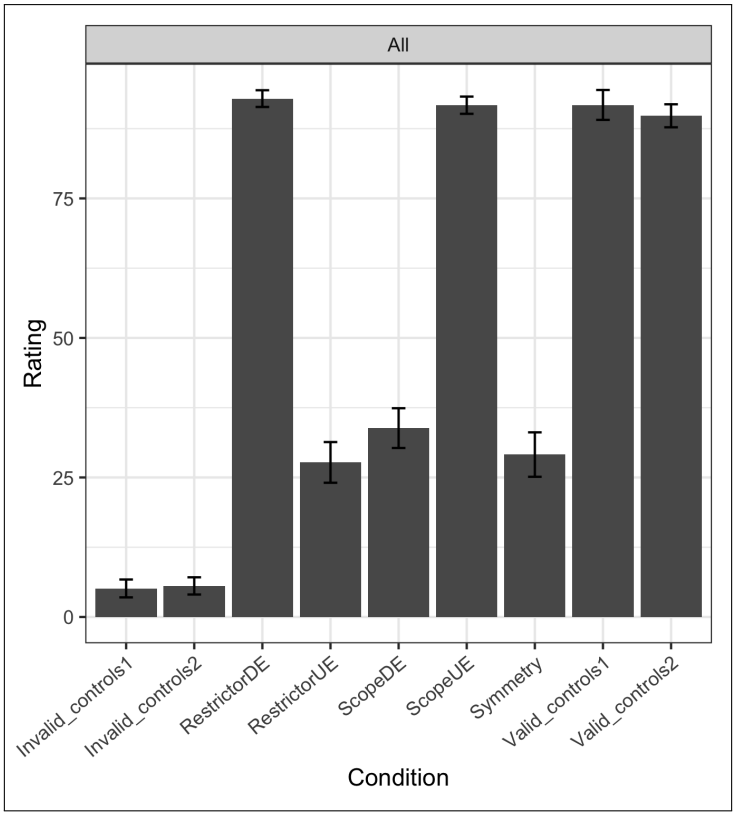


Figure 7: Results of the inference judgment task of Experiment 3 per CONDITION (Experiment 3B). Error bars represent standard errors.

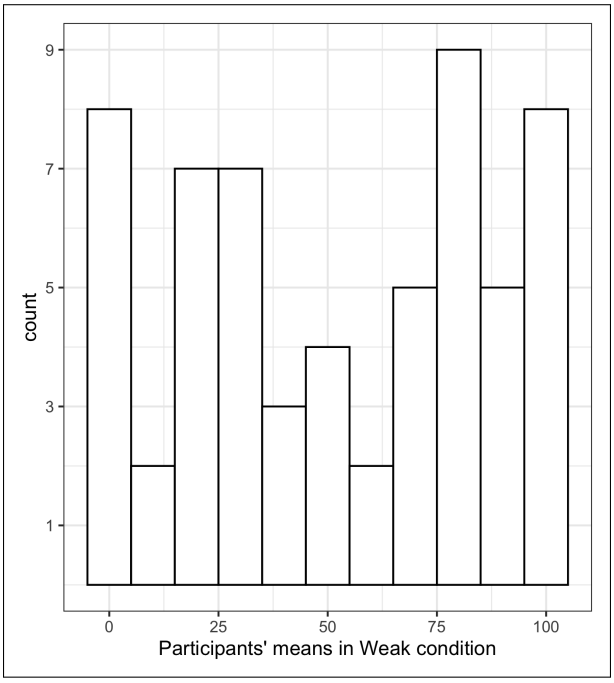


Figure 8: Exp. 3: Distribution of participants' mean responses in WEAK

The first question addressed by Experiment 3B is whether each participant's subjective perception of the monotonicity of the quantifier *all* (to be computed based on their responses in the inference judgment task) predicts the rate at which that participant

accesses the ∀-reading of donkey anaphora (to be computed based on their responses in the truth value judgment task).

According to Kanazawa's generalization, sentences headed by ↑MON↑ or ↓MON↓ have the ∃-reading as default, sentences headed by ↑MON↓ or ↓MON↑ have the ∀-reading as default reading. A way to restate Kanazawa's generalization to allow for variability in the access of ∀- and ∃-reading is in (31):

(31) *Restating Kanazawa's generalization:* The larger the difference in monotonicity properties between the restrictor and the scope of a quantifier, the more easily that quantifier receives the ∀-reading of donkey anaphora.

According to this generalization, if preservation of the subjective inferential properties of a quantifier matters for donkey anaphora interpretation, the larger the perceived difference in monotonicity between the restrictor and the scope of the quantifier *all*, the more one should access the ∀-reading of donkey anaphora in sentences headed by *all*.

How to measure the extent to which a participant accesses the ∀-reading? We assume that the amount of ∀-readings correlates negatively with the rating given to WEAK in the truth value judgment task, in which the ∀-reading is not verified: the more one accesses the ∀-reading, the lower rating they should assign to the items in WEAK.

We normalized for each participant p their responses r on items in WEAK as in (32), where $\text{mean}_p(\textsc{Strong})$ and $\text{mean}_p(\textsc{False})$ represent respectively the mean response in the STRONG and FALSE conditions of participant p. This scales the participant's responses in WEAK within their extreme judgments of STRONG and FALSE.[18]

(32)

$$\text{normalized } r = \frac{r - \text{mean}_p(\textsc{False})}{\text{mean}_p(\textsc{Strong}) - \text{mean}_p(\textsc{False})}$$

After the normalization, we excluded all responses x given to the items in WEAK such that $x < -0.5$ or $x > 1.5$; these are the items in WEAK that are judged significantly better than items in STRONG, or items that are judged significantly worse than items in the FALSE[19], and are thus likely to be errors (18 out of 360 responses).

In order to calculate the measure of subjective perception of monotonicity, we proceeded as follows. First, we ensured that the four monotonicity-related conditions (SCOPE-DE, SCOPE-UE, RESTRICTOR-DE, RESTRICTOR-UE) receive a uniform directional interpretation: responses in SCOPE-DE and RESTRICTOR-DE were kept untransformed, but responses in SCOPE-UE and RESTRICTOR-UE were reversed (a response x in the latter two conditions would become $100\% - x$). This transformation aligns the responses across the four conditions in the following sense: it measures to what extent downward entailing inferences follow, and to what extent upward entailing inferences do not follow. We refer to these as directional responses, and to the four conditions post-transformation of responses as DIR-SCOPE-DE, DIR-SCOPE-UE, DIR-RESTRICTOR-DE, DIR-RESTRICTOR-UE. Second, we normalized each participant's responses in these four conditions by scaling them within that participant's extreme judgments of VALID CONTROLS 1 and 2 on the one hand, and INVALID CONTROLS 1 and 2 on the other hand.

[18]The normalization was done for the truth value judgment task and for the inference judgment task in order to correct for different uses of response scale among the participants between the two tasks.

[19]Note that if the normalized response equals 0, this indicates that this response's rating equals the participant's average rating in the FALSE condition; if the normalized response equals 1, this indicates that this response's rating equals the participant's average rating in the STRONG condition.

The normalization procedure in the inference judgment task was completely parallel to that in the truth value judgment task. Third, using normalized responses, we computed a score that represents the subjective perception of the monotonicity of the scope for each participant as in (33a), and a score that represents the subjective perception of the monotonicity of the restrictor as in (33b), with mean_p(DIR-SCOPE-DE) being the mean of (normalized) directional responses of a participant p in the condition SCOPE-DE (and likewise for SCOPE-UE, RESTRICTOR-DE, RESTRICTOR-UE). Finally, each participant's *monotonicity index* was calculated as the absolute value of the difference between the monotonicity of the scope, calculated as in (33a), and the monotonicity of the restrictor, calculated as in (33b).

(33) a.

$$\frac{\text{mean}_p(\text{DIR-SCOPE-DE}) + \text{mean}_p(\text{DIR-SCOPE-UE})}{2}$$

b.

$$\frac{\text{mean}_p(\text{DIR-RESTRICTOR-DE}) + \text{mean}_p(\text{DIR-RESTRICTOR-UE})}{2}$$

To test whether the monotonicity index predicts the incidence of the ∀-reading of donkey anaphora, the data was subsetted to items in WEAK. A linear mixed effect model was fitted on the normalized responses to the items in WEAK, with MONOTONICITY INDEX as a fixed effect, and random by-participant intercepts. A comparison of this model with a reduced model without MONOTONICITY INDEX as a fixed effect revealed no significant effect of MONOTONICITY INDEX ($\chi^2(1) = 0.005, p = .94$). In other words, there is no evidence that subjective perception of monotonicity predicts the amount of ∀-readings of donkey anaphora.

The monotonicity index defined above is motivated by Kanazawa's generalization, according to which the difference in monotonicity between the restrictor and the scope of a quantifier mediates donkey anaphora interpretation. One may wonder, however, whether the monotonicity index is a relevant aggregate of the judgments of the four monotonicity-related conditions. For this reason, we conduct a post-hoc analysis which examines if there is an effect of subjective mononotonicity as assessed in any of the monotonicity-related conditions or linear combinations thereof on the incidence of the ∀-reading. To this end, a linear mixed effect model was fitted on the normalized responses to the items in WEAK, with each participant's mean score on normalized responses in each of DIR-SCOPE-DE, DIR-SCOPE-UE, DIR-RESTRICTOR-DE and DIR-RESTRICTOR-UE as fixed effects, and random by-participant intercepts. A comparison of this model with a random-effects only model revealed no significant difference between the two models ($\chi^2(4) = 1.11, p = .89$). In other words, there is no evidence that subjective perception of monotonicity in any of the four monotoniticy-related conditions or linear combinations thereof predicts the amount of ∀-readings of donkey anaphora.

A.6 Subjective symmetry and the amount of ∀-reading

The second question addressed by Experiment 3B is whether each participant's subjective perception of the symmetry of the quantifier *all* (to be computed based on their responses in the inference judgment task) predicts the rate at which that participant accesses the ∀-reading of donkey anaphora.

We have discussed the connection between the symmetry of the quantifier and the ∃-reading (cf. Kanazawa 1994). From this connection it follows that if preservation of symmetry-based subjective inferences matters for donkey anaphora, the less symmetric the quantifier *all* is perceived to be, the more ∀-readings of donkey anaphora it should give rise to.

We have already established how to evaluate the extent to which each participant accesses the ∀-reading in Section A.5. In order to calculate their *symmetry indices*, we normalized each participant's responses in SYMMETRY by scaling them within that participant's extreme judgments of VALID CONTROLS 1 and 2 on the one hand and INVALID CONTROLS 1 and 2 on the other hand. The normalization procedure here is completely parallel to the normalization procedure performed on responses in the four monotonicity-related conditions, and on the responses in WEAK in the truth value judgment task (cf. Section A.5). Each participant's *symmetry index* is calculated as this participant's mean of the normalized responses in SYMMETRY.

To test whether the symmetry indices predict the robustness of the ∀-reading of donkey anaphora, the data was subsetted to the items in WEAK. A linear mixed model was fitted on the normalized responses in WEAK with SYMMETRY INDEX as a fixed effect, and random by-participant intercepts. A comparison of this model with a reduced model without SYMMETRY INDEX as a fixed effect revealed no significant effect of SYMMETRY INDEX ($\chi^2(1) = 0.6, p = .43$). In other words, there is no evidence that subjective perception of symmetry predicts the amount of ∀-readings of donkey anaphora.

A.7 Discussion of Experiment 3B

In Experiment 3B, we investigated whether subjective perception of the monotonicity and symmetry of the quantifier *all* predict the interpretation of donkey anaphora with this quantifier. We found considerable between-participant variation in the amount of ∀-readings among different participants (cf. Figure 8), but this variation does not correlate with the participants' subjective monotonicity and symmetry of the quantifier *all*. This fact is a challenge for a specific instantiation of Kanazawa's (1994) theory according to which the interpretation of donkey anaphora in the scope of a quantifier is selected so that it preserves subjective inferential patterns based on the monotonicity and/or symmetry of the quantifier in question.

References

Adrian Brasoveanu. *Structured Nominal and Modal Reference*. PhD thesis, Rutgers University, 2007.

Lucas Champollion, Dylan Bumford, and Robert Henderson. Donkeys under discussion. *Semantics and Pragmatics*, 12, 2019.

Emmanuel Chemla, Vincent Homer, and Daniel Rothschild. Modularity and intuitions in formal semantics: the case of polarity items. *Linguistics and Philosophy*, 34(6): 537–570, 2011.

Gennaro Chierchia. *Dynamics of Meaning: Anaphora, Presupposition, and the Theory of Grammar*. University of Chicago Press, Chicago, 1995.

Robin Cooper. The interpretation of pronouns. In Frank Heny and Helmut Schnelle, editors, *Syntax and Semantics 10: Selections from the Third Groeningen Round Table*, pages 61–92. Academic Press, New York, 1979.

Francesca Foppolo. The puzzle of donkey anaphora resolution. In *Proceedings of NELS*, 2008.

Bart Geurts. Donkey business. *Linguistics and Philosophy*, 25(2):129–156, 2002. doi: 10.1023/A:1014624331450.

Irene Heim. *The Semantics of Definite and Indefinite Noun Phrases*. PhD thesis, University of Massachusetts, Amherst, 1982.

Irene Heim. E-type pronouns and donkey anaphora. *Linguistics and Philosophy*, 13(2): 137–177, 1990.

Nirit Kadmon. *On unique and non-unique reference and asymmetric quantification*. PhD thesis, University of Massachusetts, Amherst, 1987.

Nirit Kadmon. Uniqueness. *Linguistics and Philosophy*, 13(3):273–324, 1990.

Makoto Kanazawa. Weak vs. strong readings of donkey sentences and monotonicity inferences in a dynamic setting. *Linguistics and Philosophy*, 17(2):109–158, 1994. doi: 10.1007/BF00984775.

Manfred Krifka. Pragmatic strengthening in plural predications and donkey sentences. In *Proceedings of SALT 6*, pages 136–153, 1996.

Reinhard Muskens. Combining Montague Semantics and Discourse Representation. *Linguistics and Philosophy*, 19(2):143–186, 1996.

Stanley Peters and Dag Westerstål. *Quantifiers in Language and Logic*. Oxford University Press, Oxford, 2006.

Craige Roberts. Information structure: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5:1–69, 2012.

Mats Rooth. Noun phrase interpretation in Montague Grammar, File Change Semantics, and Situation Semantics. In Peter Gärdenfors, editor, *Generalized Quantifiers: Linguistic and Logical Approaches*, pages 237–268. Reidel, Dordrecht, 1987.

Mats Rooth. A theory of focus interpretation. *Natural language semantics*, 1(1):75–116, 1992.

Lenhart K. Schubert and Francis Jeffery Pelletier. Generically speaking, or, using discourse representation theory to interpret generics. In Gennaro Chierchia, Barbara H. Partee, and Raymond Turner, editors, *Properties, Types and Meaning II: Semantic Issues*, pages 193–268. Kluwer Academic Publishers, 1989.

Chao Sun, Daniel Rothschild, and Richard Breheny. Exploring the existential/universal ambiguity in singular donkey sentences. In *Proceedings of Sinn und Bedeutung 24*, volume 2, pages 289–305, 2020. doi: https://doi.org/10.18148/sub/2020.v24i2.898.

Arnim von Stechow. Focusing and backgrounding operators. *Discourse particles*, 6:37–84, 1991.

Youngeun Yoon. *Weak and Strong Interpretations of Quantifiers and Definite NPs in English and Korean*. PhD thesis, University of Texas at Austin, 1994.

Youngeun Yoon. Total and partial predicates and the weak and strong interpretations. *Natural Language Semantics*, 4(3):217–236, 1996. doi: 10.1007/BF00372820.