Donkey Anaphora in Non-Monotonic Environments*

Milica Denić
Universiteit van Amsterdam

Yasutada Sudo
University College London

July 29, 2020

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 $\exists$-reading and the $\forall$-reading (Brasoveanu 2007, Champollion, Bumford, and Henderson 2018, Chierchia 1995, Cooper 1979, Foppolo 2008, Geurts 2002, Kanazawa 1994, Krifka 1996, Rooth 1987, Schubert and Pelletier 1989, Sun, Breheny, and Rothschild 2019, Yoon 1994, 1996).\(^2\) For instance, (1) is most naturally interpreted with a $\forall$-reading, (2a), but as Chierchia (1995) points out, when a context like (3) is provided, a $\exists$-reading, (2b), becomes available.

(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.

(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

---

\(^1\)We wish to thank Emmanuel Chemla for many important discussions and suggestions. We also thank Keny Chatain, Gennaro Chierchia, 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.

\(^2\)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.

---

2For some speakers (1) is associated with a uniqueness presupposition that every farmer who owns a donkey owns only one, and with this presupposition the $\exists$ and $\forall$ 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. (2018) for related discussion.
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)

∃-readings are also observed with examples that make the ∀-reading implausible to be true, as in (4) (Chierchia 1995, Kanazawa 1994, Schubert and Pelletier 1989).

(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)

In addition, which quantifier is used affects which reading of a donkey pronoun is more prominent (Chierchia 1995, Kanazawa 1994, Yoon 1994, 1996, Champollion et al. 2018). Generally, donkey pronouns in the scope of universal quantifiers (e.g. every, all, each, etc.) tend to receive ∀-readings, while those 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 these determiners cannot receive ∀-readings (Kanazawa 1994, Chierchia 1995, Geurts 2002). Chierchia (1995:p. 65) points out that (7) is most naturally read under the ∀-reading (see Yoon 1996:p. 229 for more examples).

(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. 2019 for more than two).

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

Thus, across sentences and contexts, one may find both ∀- and ∃-readings for each of the above quantifiers, but it is also considered that ∀-readings are more commonly observed with quantifiers like every and not every than with quantifiers like no and some, which seem to more commonly receive ∃-readings. These patterns have been corroborated by experimental studies (Yoon 1994, 1996, Geurts 2002, Foppolo 2008, Sun et al. 2019). Therefore, there seem to be at least two classes of quantifiers with respect to which of the ∀- and ∃-construals they more commonly receive across sentences and contexts. We will speak of these as the ‘default-readings’ of these quantifiers. It should be kept in mind, however, that for each sentence (in a specific context), the default reading of the sentence (in an intuitive sentence of the term) may be the opposite one, e.g. the default reading of the quantifier no is the ∃-reading, but the default reading of a sentence like

Chierchia (1995) attributes (ia) 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.

Chierchia (1995) cites a manuscript version of Kanazawa (1994) for this example, but its published version does not contain it.
(7) is arguably the ∀-reading. Our main interest in this paper is the default readings of quantifiers in donkey sentences, rather than the default readings of donkey sentences themselves.

Donkey anaphora involving monotonic quantifiers like the examples we have seen so far is relatively well studied. In particular, Kanazawa (1994) proposes the following generalization.

(9) a. Default readings for ↑MON↑ 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. 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 who own a donkey beat it.

Kanazawa (1994) offers some conjectures about why this might be so, which we will discuss in some detail. More recently, Champollion et al. (2018) put forward an alternative theory to Kanazawa’s (1994) that predicts a ‘conjunctive reading’ for sentences like (10) (modulo contextual factors), which is the conjunction of the ∃- and ∀-readings. Thus, under this view, (10) is purported to be true if and only if there are exactly three donkey-owning farmers who beat their donkeys, and they all beat all of their donkeys.

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.

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. (2018) 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 will discuss theoretical implications of our experimental findings with respect to Kanazawa’s (1994) and Champollion et al.’s (2018) proposals. Section 6 contains conclusions and further directions. The experimental data, the R script used for analysis, and the design files can be found at XXX.

The literature sometimes mentions proportional quantifiers like most, which are non-monotonic with respect to the NP argument. We will come back to it in Section 5.1.
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) summarizes his generalization in terms of monotonicity as in (9), and proposes that this generalization can be understood 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 ↓MON↓, and in a donkey sentence, the ∃-reading, but not the ∀-reading, preserves this monotonicity profile with donkey anaphora. To see this, consider (11).

\begin{enumerate}
  \item No farmer who owns a donkey beats it.
  \item No farmer who owns a young donkey beats it.
\end{enumerate}

If both of these sentences received ∀-readings, then (11a) would not entail (11b). That is, (11a) under the ∀-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 ∀-reading of (11a) would be true and the ∀-reading of (11b) would be false. On the other hand, the entailment from (11a) to (11b) would go through under the ∃-readings of the sentences. Consequently, under Kanazawa’s (1994) proposal, the default reading of no in donkey sentences is the ∃-reading.

Let us look at another example, this time with a universal quantifier, which is ↓MON↑.

\begin{enumerate}
  \item Every farmer who owns a donkey beats it.
  \item Every farmer who owns a young donkey beats it.
\end{enumerate}

First consider the ∃-readings of these sentences, under which (12a) would not entail (12b). Specifically, in the following situation the ∃-reading of (12a) is true while the ∃-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 ∀-readings. Therefore, according to Kanazawa (1994), the default reading of every is the ∀-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 two seem to prefer the ∃-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.

\begin{enumerate}[start=13]
  \item A quantifier $Q$ is left-continuous if and only if for all $A, B, C, X \subseteq D_e$ such that $A \subseteq B \subseteq C$, $Q(A)(X)$ and $Q(C)(X)$ together imply $Q(B)(X)$.
  \item A quantifier $Q$ is symmetric if and only if for all $A, B \subseteq D_e$, $Q(A)(B) = Q(B)(A)$.
\end{enumerate}

The classical generalized quantifier corresponding to exactly three is both left-continuous and symmetric.\(^6\) As Kanazawa (1994) observes, in donkey anaphora, the ∃-reading pre-
serves both of these properties, but not the $\forall$-reading. Let us see this first for the case of left-continuity. Consider the following sentences.

(15) a. Exactly three farmers who own livestock beat it.
    b. Exactly three farmers who own a donkey beat it.
    c. Exactly three farmers who own a young donkey beat it.

Take the following situation, which makes (15a) and (15c) true under the $\forall$-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 $\forall$-reading. On the other hand, under the $\exists$-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 livestock.

Let us now turn to symmetry. One issue is 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 if it is intersective (e.g. Peters and Westerståhl 2006), i.e. for all $A, B \subseteq D_e$, $Q(A)(B) = Q(A \cap B)(D_e)$, which can be checked without disrupting donkey anaphora.\footnote{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 conservativity, $Q(A \cap B)(A \cap B)$ is also equivalent, which is obviously symmetric.

Let us now 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 $\exists$-reading of (16a) is equivalent to (16b) but its $\forall$-reading is not. Since we only discuss conservative quantifiers, we will not distinguish symmetry and intersectivity below.

Kanazawa (1994) proposes that if at least one of left-continuity and symmetry needs to be preserved, then the preference for $\exists$-readings for quantifiers like exactly three could be accounted for, but he does not provide a definitive answer as to whether or not both of them matter, 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. 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 $|B \cap X| \geq 3$, and if $B \subseteq C$, then $|B \cap X| \leq 3$. Thus if $A \subseteq B \subseteq C$, then $|B \cap X| = 3$. \qed
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.\(^8\) 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, contrary to what Kanazawa (1994) claims. On the other hand, if that is not the case, then monotonicity preservation will be crucial, as there are monotonic but non-symemtric quantifiers like every, whose default readings cannot be explained in terms of symmetry preservation alone.

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, but 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.\(^9\) Specifically, according to the first hypothesis, all but one has no property to preserve, so neither reading should be preferred. According to the other two hypotheses, its left-continuity needs to be preserved.

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).

\[(17) \quad \begin{align*}
a. \text{ All but one of the farmers who own livestock beat it.} \\
b. \text{ All but one of the farmers who own a donkey beat it.} \\
c. \text{ All but one of the farmers who own a young donkey beat it.}
\end{align*}\]

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 livestock; All the other farmers beat at least some of their livestock; Farmer B owns some young

---

\(^8\)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 \(B\) 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 \(B\) such that \(A \subseteq B \subseteq C\).

\(^9\)The non-monotonicity of all but one is obvious. 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 \(\llbracket \text{all but one} \rrbracket (A)(B)\) if \(|A| - |A \cap B| = 1:\)

**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 \(|B \cap X| \geq n - 1\), and if \(B \subseteq C\), then \(|B \cap X| \leq n - 1\). Thus if \(A \subseteq B \subseteq C\), then \(|B \cap X| = n - 1\).
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 (17c) together entail the $\forall$ reading of (17b).

Therefore, if it turns out that the preferred reading of all but one is the $\forall$-reading, then we can reject the first hypothesis; on the other hand, if it turns out that neither reading is preferred, we can reject the latter two. Table 1 summarizes the predictions of the three hypotheses.

<table>
<thead>
<tr>
<th></th>
<th>exactly three</th>
<th>all but one</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. symmetry + monotonicity</td>
<td>$\exists$</td>
<td>No preference</td>
</tr>
<tr>
<td>2. left-continuity (+ monotonicity)</td>
<td>$\exists$</td>
<td>$\forall$</td>
</tr>
<tr>
<td>3. symmetry + left-continuity (+ monotonicity)</td>
<td>$\exists$</td>
<td>$\forall$</td>
</tr>
</tbody>
</table>

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 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 $\exists$-reading in both components, (ii) a $\forall$-reading in both components, (iii) an $\exists$-reading in the upward component combined with a $\forall$-reading in the downward component and (iv) a $\forall$-reading in the upward component combined with an $\exists$-reading in the downward component. We will label these readings as (i) $\exists \exists$, (ii) $\forall \forall$, (iii) $\exists \forall$, (iv) $\forall \exists$, 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.

- $\exists \exists$: At least three squares that are above a heart(s) are connected to some of those hearts and at most three squares that are above a heart(s) are connected to some of those hearts.
- $\forall \forall$: At least three squares that are above a heart(s) are connected to all of those hearts and at most three squares that are above a heart(s) are connected to all of those hearts.
- $\exists \forall$: At least three squares that are above a heart(s) are connected to all of those hearts and at most three squares that are above a heart(s) are connected to some of those hearts.
- $\forall \exists$: At least three squares that are above a heart(s) are connected to some of those hearts and at most three squares that are above a heart(s) 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](image)

It is important to note that \( \exists \) corresponds to what is typically referred to as the \( \exists \)-reading of donkey anaphora; \( \forall \) corresponds to the \( \forall \)-reading; \( \exists \) \( \forall \) corresponds to the ‘conjunctive reading’ discussed by Champollion et al.’s (2018) (i.e. conjunction of \( \exists \) and \( \forall \) reading). On the other hand, \( \exists \) \( \forall \) 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 a 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 \( \exists \) \( \forall \) is true, e.g. Figure 2a.
• DEweak-UEstrong: Only $\forall$ and $\exists$ are true, e.g. Figure 2b.

• DEstrong-UEweak: Only $\exists$ and $\exists$ 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 $\forall$-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 $\exists$-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). Each of the six conditions had six items, there were thus 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.
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 they 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.

![Figure 3: Results of the two experiments per condition. Error bars represent standard errors.](image)

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

No evidence for $\frac{3}{9}$. If the weakest reading, $\frac{3}{9}$, has been accessed, DEWEAK-UEWEAK, which validates $\frac{3}{9}$, should receive higher rating than the control conditions, which validate none of the readings. The data was subverted to the items in DEFALSE-UESTRONG and
### Table 2: Target conditions and the readings that they make true.

<table>
<thead>
<tr>
<th>Condition</th>
<th>( \frac{2}{3} )</th>
<th>( \frac{3}{4} )</th>
<th>( \frac{4}{5} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{DEstrong-UEstrong} )</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>( \text{DEstrong-UEweak} )</td>
<td>F</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>( \text{DEweak-UEstrong} )</td>
<td>F</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>( \text{DEweak-UEweak} )</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>( \text{DEfalse-UEstrong} )</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>( \text{DEstrong-UEfalse} )</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th>Condition</th>
<th>Exp. 1: Exactly three</th>
<th>Exp. 2: All but one</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{DEweak-UEweak} )</td>
<td>13.4 (3.1)</td>
<td>12.4 (2.6)</td>
</tr>
<tr>
<td>( \text{DEweak-UEstrong} )</td>
<td>13.2 (3.1)</td>
<td>25.3 (4.3)</td>
</tr>
<tr>
<td>( \text{DEstrong-UEweak} )</td>
<td>85.2 (2.8)</td>
<td>67 (4.3)</td>
</tr>
<tr>
<td>( \text{DEstrong-UEstrong} )</td>
<td>86.7 (2.7)</td>
<td>82.3 (3.2)</td>
</tr>
<tr>
<td>( \text{DEfalse-UEstrong} )</td>
<td>13 (2.8)</td>
<td>9.6 (2.3)</td>
</tr>
<tr>
<td>( \text{DEstrong-UEfalse} )</td>
<td>6.5 (2)</td>
<td>8.1 (2.2)</td>
</tr>
</tbody>
</table>

DE\text{weak-UEweak} conditions\textsuperscript{10}. 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 \((\chi(1) = 0.06, p = .8)\). There is thus no evidence for the existence of \( \frac{3}{4} \) reading with exactly three.

**No evidence for \( \frac{2}{3} \)** If \( \frac{2}{3} \) has been accessed, DE\text{weak-UEstrong}, which validates both \( \frac{3}{4} \) and \( \frac{2}{3} \), should receive higher rating than DE\text{weak-UEweak}, which only validates \( \frac{2}{3} \). The data was subsetted to the items in DE\text{weak-UEstrong} and DE\text{weak-UEweak} conditions. 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 \((\chi(1) = 0.02, p = .9)\). There is thus no evidence for the existence of \( \frac{2}{3} \) reading with exactly three.

**Evidence for \( \frac{3}{4} \)** If \( \frac{3}{4} \) has been accessed, DE\text{strong-UEweak}, which validates both \( \frac{3}{4} \) and \( \frac{3}{5} \), should receive higher rating than DE\text{weak-UEweak}, which only validates \( \frac{3}{5} \). The data was subsetted to items in DE\text{strong-UEweak} and DE\text{weak-UEweak} conditions. 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 \((\chi(1) = 0.02, p = .9)\).

\textsuperscript{10}DE\text{false-UEstrong} was chosen rather than DE\text{strong-UEfalse} because the mean rating of DE\text{false-UEstrong} was higher than that of DE\text{strong-UEfalse}, and thus provides a stricter requirement for the detection of \( \frac{2}{3} \).
condition \((\chi(1) = 100, p < .001)\). Our results thus provide evidence for the existence of \(\frac{3}{3}\) reading with exactly three.

No evidence for \(\frac{7}{7}\). In order to uncover whether the \(\frac{7}{7}\) reading is available or not, we cannot simply compare DEstrong-UEstrong condition, which is the only condition validating \(\frac{7}{7}\), to some other condition. The reason is the following. Suppose the reading \(\frac{7}{7}\) is never accessed, while the other three readings (i.e., \(\frac{5}{5}, \frac{3}{3},\) and \(\frac{3}{7}\)) 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 condition 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 \(\frac{7}{7}\).

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 \(\frac{5}{5}\) and \(\frac{7}{7}\). In other words, they (at most) accessed either (i) \(\frac{5}{5}, \frac{3}{7}\), and \(\frac{3}{7}\) or (ii) \(\frac{3}{3}, \frac{3}{7},\) and \(\frac{3}{7}\). This further means that, for the participants who did not access \(\frac{7}{7}\), the only reading which is true in DEstrong-UEstrong but not in DEstrong-UEweak is \(\frac{7}{7}\). Likewise, for participants who did not access \(\frac{7}{7}\), the only reading which is true in DEstrong-UEstrong but not in DEweak-UEstrong is \(\frac{7}{7}\). 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 \(\frac{7}{7}\). 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 \((\chi(1) = 0.01, p = .9)\). 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 \((SD = 2.8)\), while their mean rating in DEstrong-UEstrong was 88.7 \((SD = 3.4)\). There is thus no evidence for the existence of \(\frac{7}{7}\) with exactly three.

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

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, any differences between the results of the two experiments have to be due to the linguistic, rather than 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.

**No evidence for** \( \text{D} \)

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 \( \text{D} \) with all but one.

**Evidence for** \( \text{V} \)

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 \( \text{V} \) with all but one.

**Evidence for** \( \text{D} \)

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 \( \text{D} \) with all but one.

**Weak evidence for** \( \text{V} \)

As in Experiment 1, in order to determine whether participants have accessed \( \text{V} \), 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 a borderline effect of condition (DEstrong-UEstrong vs. other) \((\chi(1) = 3.62, p = .057)\). 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 \((SD = 5.6)\), while the mean rating in DEstrong-UEstrong was 82.3 \((SD = 4.6)\). This suggests that \( \text{V} \) might be available with all but one, but further research is needed to establish this conclusively.
Summarizing the results of Experiment 2, we detected \( \frac{3}{3} \) (i.e. \( \exists \)-reading) and \( \frac{4}{3} \) (i.e. \( \forall \)-reading) reading of donkey anaphora in the scope of all but one. In addition, we found borderline evidence for the existence of \( \frac{3}{3} \). Focusing on the two readings for whose existence we have clear evidence, \( \exists \)-reading is nonetheless preferred to \( \forall \)-reading, as DESTRONG-UEWEAK condition is judged significantly better than DEWEAK-UESTRONG condition, as evidenced by analyses parallel to those reported above (\( \chi(1) = 26.8, p < .001 \)).

5 General discussion

Let us summarize the main empirical findings from our two experiments.

1. We obtained no evidence that \( \frac{3}{3} \) is available with either quantifier. This is unsurprising; this reading, even though logically possible, has not been reported in the literature so far.

2. There is clear evidence that \( \frac{3}{3} \) 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. Most interestingly, we find differences between all but one and exactly three with respect to the availability of \( \frac{4}{3} \).

4. Finally, we find suggestive evidence that \( \frac{4}{3} \) is available with all but one, but further data need be assessed before a firm conclusion could be drawn. No such evidence is found for exactly three.

Let us now address a potential reproach to the generalizability of these findings. The fact that only \( \frac{3}{3} \)-reading was detected with exactly three and that \( \frac{3}{3} \)-reading 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 \( \frac{3}{3} \)-reading (i.e. \( \exists \)-reading), and thus masking the existence of other readings, such as for instance \( \forall \)-reading with exactly three and \( \frac{4}{3} \)-reading with both quantifiers.

Experiments 1 and 2 were designed in such a way as to allow comparison with Experiment 3 reported in Appendix, which investigated donkey anaphora interpretation with quantifier all, and allows us to evaluate whether the specifics of our experimental setup are to blame for the undetected readings. 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 the truth value judgment task of Experiment 3, of which six validated both \( \forall \)- and \( \exists \)-reading of donkey anaphora, six validated only the \( \exists \)-reading, and six didn’t validate either. (The comparability stems from the fact that, modulo \( \frac{3}{4} \) and \( \frac{4}{4} \), DESTRONG-UESTRONG validates both \( \forall \)- and \( \exists \)-reading of donkey anaphora, DESTRONG-UEWEAK validates only the \( \exists \)-reading, and the false controls do not validate either.) If the specifics of the experimental setup are biasing strongly towards the \( \exists \)-reading in Experiments 1 and 2, we expect to observe this bias in Experiment 3 as well. As is evident from the results of Experiment 3 with quantifier all reported in Appendix (cf. Figure 5a), the condition which makes the \( \exists \)-reading but not the \( \forall \)-reading true is judged significantly lower in Experiment 3 than in Experiments 1 and 2. This demonstrates that the low rates of readings other than the \( \exists \)-reading with all but one and exactly three are due to those specific quantifiers rather than due to the experimental setup.
In light of these observations, we will now discuss potential theoretical implications.

5.1 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 \textit{exactly three} is the \(\exists\)-reading, which we’ve been calling \(\frac{\exists}{3}\).

- If left-continuity needs to be preserved, the default reading of \textit{all but one} should be the \(\forall\)-reading, which we called \(\frac{\forall}{5}\); if it need not be preserved, then there shouldn’t be a preference.

Our experimental results suggest that \textit{all but one} can have a \(\forall\)-reading, while we observed no evidence for the \(\forall\)-reading for \textit{exactly three}. This is consistent with the first point above. What about the second point? That the \(\exists\)-reading is preferred for \textit{all but one} is more in line with the hypothesis where neither \(\exists\) nor \(\forall\) reading is default. This of course depends on the baseline rate for the \(\exists\)-reading, but as mentioned above, in Experiment 3 reported in Appendix, \textit{all} received the \(\forall\)-reading at a substantially higher rate. We take this as evidence that left-continuity is not required to be preserved. 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.

In order to further buttress this conclusion, let us look at a non-left-continuous by symmetric quantifier, \textit{exactly three or exactly five}. It’s symmetry is obvious. To see the lack of left-continuity of \textit{exactly three or exactly five}, consider (21).

\begin{enumerate}
\item a. Exactly three or exactly five Europeans are funny.
\item b. Exactly three or exactly five Germans are funny.
\item c. Exactly three or exactly five Berliners are funny.
\end{enumerate}

The following situation makes (21a) and (21c) true but (21b) false: There are exactly three funny Berliners, exactly one funny German living in Stuttgart, and exactly one funny Italian; No other Europeans are funny.

According to the symmetry + monotonicity hypothesis, the default reading for this quantifier should be the \(\exists\)-reading, just like for \textit{exactly three}. On the other hand, according to the hypothesis according to which only left-continuity needs to be preserved, \textit{exactly three or exactly five} should have no preference, while \textit{exactly three} should prefer the \(\exists\)-reading.

\begin{enumerate}
\item a. Exactly three farmers who own a donkey beat it.
\item b. Exactly three or exactly five farmers who own a donkey beat it.
\end{enumerate}

We have no quantitative data at this point, but it appears to us that these two quantifiers do not show marked difference in donkey sentences, such supports the view that symmetry is one of the properties to be preserved.

5.2 The Conjunctive Reading and Pragmatic Factors

We now turn to a recent theory proposed by Champollion et al. (2018) that makes predictions about which readings of donkey anaphora are preferred under which quantifiers.

16
According to them, the default reading, which we refer to as ‘conjunctive reading’, for each case is true if both the $\exists$- and $\forall$-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:\footnote{Champollion et al. (2018) 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 $\exists$-readings.}

\begin{enumerate}
\item \textbf{Every farmer who owns a donkey beats it} is:
\begin{enumerate}
\item \textit{true}, if every donkey-owning farmer beats all of their donkeys;
\item \textit{false}, if at least one donkey-owning farmer beats none of their donkeys; and
\item \textit{neither true nor false}, otherwise.
\end{enumerate}
\item \textbf{No farmer who owns a donkey beats it} is:
\begin{enumerate}
\item \textit{true}, if no donkey-owning farmer beats any of their donkeys;
\item \textit{false}, if some donkey-owning farmer beats all of their donkeys; and
\item \textit{neither true nor false}, otherwise.
\end{enumerate}
\end{enumerate}

In addition to this semantics, Champollion et al. (2018) assume that depending on various contextual factors (such as Questions under Discussion), 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 collapsed into false, the reading amounts to the $\forall$-reading in (23) and the $\exists$-reading in (24) in a bivalent setting. Alternatively, when the false and neither true nor false cases are collapsed into true, the reading amounts to the $\exists$-reading in (23) and the $\forall$-reading in (24).

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

\begin{enumerate}
\item \textbf{Exactly three farmers who own a donkey beat it} is:
\begin{enumerate}
\item \textit{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;
\item \textit{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
\item \textit{neither true nor false}, otherwise.
\end{enumerate}
\item \textbf{All but one farmer who owns a donkey beats it} is:
\begin{enumerate}
\item \textit{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;
\item \textit{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
\item \textit{neither true nor false}, otherwise.
\end{enumerate}
\end{enumerate}

To put it differently, for both of these cases, the sentence is true when $\forall_3$ is true, and it is false when neither $\exists_3$ nor $\forall_3$ is true, and in cases where $\exists_3$ or $\forall_3$ is true but not $\forall_3$, the sentence is semantically neither true nor false. If the \textit{neither true nor false} cases are collapsed into false, the reading basically amounts to $\forall_3$.\footnote{Champollion et al. (2018) 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 $\exists$-readings.}
The results of our experiment pose several challenges for this theory. Firstly, as we have discussed, it is not very easy to give empirical evidence for the availability of the strongest reading, \(\exists\), for sentences with donkey anaphora headed with exactly three or all but one. Of course, this does not in itself disprove Champollion et al.’s (2018) analysis, but it should be ideally given convincing evidence that this reading is ever accessible.

Secondly, we obtained evidence for \(\forall\)-readings for all but one but not for exactly three. For exactly three, the only reading detected in our experiment was the \(\exists\)-reading. According to Champollion et al.’s (2018) 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. (2018:§ 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, following Geurts (2002) they suggest 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 \(\forall\)- and \(\exists\)-readings simply collapse, making the conjunctive reading true, even if it is false with respect to the entire scene. For example, for (27), the relevant subpart of the visual scene will contain all the triangles but not all the hearts: it only contains those hearts that are connected to a triangle.

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

As a consequence, Champollion et al. (2018) claim, the sentence containing a symmetric quantifier is accepted often in situations where what is traditionally called the \(\exists\)-reading is true, even when the conjunctive reading is false. Furthermore, by assumption, this strategy does not apply to non-intersective quantifiers. This creates room to capture the difference between the two quantifiers. However, as far as we know, no independent evidence has been raised at this point for Geurt’s strategy, and its empirical properties are yet to be understood, without which the above explanation of our observations under Champollion et al.‘s theory remains incomplete.

Another potential way to make Champollion et al.’s (2018) theory compatible with our experimental results would be to assume that the two non-monotonic quantifiers preferred different types of Questions under Discussion in such a way that exactly three is consistently associated with Questions under Discussion that give rise to \(\exists\), while all but one is also compatible with Questions under Discussion that give rise to \(\forall\). Thus, in this case too, their theory remains incomplete without a theory of how Questions under Discussion are determined and how that interacts with the quantifier in the sentence.

6 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. 2019). Our results indicate that \(\exists\) is available with both quantifiers, while we only obtained evidence of \(\forall\) for the latter, for which suggestive evidence also exists of \(\forall\). We argued that these results support a version of Kanazawa’s (1994)’s hypothesis that symmetry needs to be preserved, together with monotonicity.
Can one provide stronger empirical support that either monotonicity or symmetry matter for the interpretation of donkey anaphora? In the aforementioned Experiment 3, which we report in Appendix, we ask whether participants’ subjective perception of the monotonicity profile and of 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 (2018). Our results pose challenges for that theory as well by calling for stronger evidence for the existence of \( \xi \), and for an addition to their theory which would predict the observed differences between all but one and exactly three.

### A Experiment on Subjective Monotonicity and Symmetry

In this appendix we report on a third experiment, whose purpose was to investigate whether the preservation of subjective inferential patterns matters for the interpretation of donkey anaphora. 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 NPI licensing. Focusing on quantifier all, we ask whether participants’ perceived monotonicity and symmetry properties of this quantifier explain the extent to which different readings of donkey anaphora are available with this quantifier.

To preview the findings, the \( \forall \)-reading is robustly accessed with all, but there seems to be no correlation with either of the subjective versions of the two logical properties. This suggests that difference in subjective monotonicity between exactly one and all but one explains the difference observed in the results of the two experiments. This suggests that preservation of inferential patterns that are due to either monotonicity or symmetry is not what’s behind reading preferences of donkey anaphora. In other words, these results suggest that, if either of these two properties matters at all, this may be for reasons other than preserving these inferential patterns.

#### A.1 Tasks

Experiment 3 had two tasks: Truth value judgment task, which was administered first, followed by Inference judgment task. 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.

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

Truth value judgment task Sentences in this task were always of the form as in (28):

\[(28)\quad \text{All of the } \langle \text{squares, triangles} \rangle \text{ that are above a } \langle \text{star, heart} \rangle \text{ are connected to it.}\]

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 4.

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

- **Weak**, in which only the \(\exists\)-reading is true, cf. Figure 4a.
- **Strong**, in which both \(\forall\)-reading and \(\exists\)-reading are true, cf. Figure 4b.

In addition, there was one control condition, (False), in which neither of the two readings were true (cf. Figure 4c). This amounts to a total of three conditions (two target and one control). 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 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).
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 $\text{All } X Y$.

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 randomized order for each participant.

For the items in Symmetry, participants had to evaluate to what extent (29a) suggests (29b) is true, with $\{X, Y\}$ taken from the list in (29c). Each of the pairs from (29c) appeared in exactly one of the six items in this condition.

(29) a. All of the $X$s from Zoopiter are $Y$s.
   b. All of the $Y$s are $X$s from Zoopiter.
   c. $\{\text{mosquito, cinema enthusiast}, \text{dove, literature enthusiast}, \text{poodle, music enthusiast}, \text{crocodile, theater enthusiast}, \text{tarantula, travel enthusiast}, \text{cobra, museum enthusiast}\}$

For the items in Restrictor-DE, participants had to evaluate to what extent (30a) suggests (30b) is true; for items in Restrictor-UE, they had to evaluate to what extent (30b) suggests (30a). For these two conditions, $\{X, Y\}$ were taken from the list in (30c), and $Z$ was taken from (30d). Each of the pairs from (30c) and each of the elements from (30d) appeared in exactly one of the six items in each of these two conditions.

(30) a. All of the $X$s from Zoopiter $Z$.
   b. All of the $Y$s from Zoopiter $Z$.
   c. $\{\text{spider, tarantula}, \text{snake, cobra}, \text{dog, poodle}, \text{insect, mosquito}, \text{bird, dove}, \text{reptile, crocodile}\}$
   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 (31a) suggests (31b) is true; for items in Scope-UE, they had to evaluate to what extent (31b) suggests (31a) is true. For these two conditions, $\{Y, Z\}$ were taken from the list in (31c), and $X$ was taken from (31d). Each of the pairs from (31c) and each of the elements from (31d) appeared in exactly one of the six items in each of these two conditions.

(31) a. All of the $X$s from Zoopiter $Y$.
   b. All of the $X$s from Zoopiter $Z$.
   c. $\{\text{planted a flower, planted a rose}, \text{watched a film, watched a documentary}, \text{wrote a book, wrote a novel}, \text{talked to an artist, talked to a sculptor}, \text{travelled to Europe, travelled to France}, \text{prepared an Asian dish, prepared a Japanese dish}\}$
   d. poodle, mosquito, dove, crocodile, tarantula, cobra

For the items in Valid Controls 1, participants had to evaluate to what extent
sentences such as (32a) suggest sentences such as (32b) are true. For the items in Valid Controls 2, participants had to evaluate to what extent sentences such as (32a) suggest sentences such as (32c) are true. For the items in Invalid Controls 1, participants had to evaluate to what extent sentences such as (32a) suggest sentences such as (32d) are true. For the items in Invalid Controls 2, participants had to evaluate to what extent sentences such as (32a) suggest sentences such as (32e) are true. Nouns appearing in the restrictor of the quantifier were taken from (31d), and predicates in the scope of the quantifier from (30d); each noun from (31d) and each predicate from (30d) appeared in exactly one of the six items in each of the four control conditions.

(32) 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

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 were 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. 60 participants were thus kept for the analysis.

### A.4 Raw results

The results obtained in the truth value judgment task are summarized in Figure 5a and Table 4, and the results obtained in the inference judgment task are summarized in Figure 5b and Table 5.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Mean rating (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong</td>
<td>92.7 (2.3)</td>
</tr>
<tr>
<td>Weak</td>
<td>52.4 (4.5)</td>
</tr>
<tr>
<td>False</td>
<td>5.5 (1.1)</td>
</tr>
</tbody>
</table>

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

### A.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.
Figure 5: Results of Experiment 3 per task per condition. Error bars represent standard errors.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Mean rating (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>INVALID CONTROLS 1</td>
<td>5.1 (1.6)</td>
</tr>
<tr>
<td>INVALID CONTROLS 2</td>
<td>5.6 (1.5)</td>
</tr>
<tr>
<td>RESTRICTOR-DE</td>
<td>92.9 (1.5)</td>
</tr>
<tr>
<td>RESTRICTOR-UE</td>
<td>27.7 (3.6)</td>
</tr>
<tr>
<td>SCOPE-DE</td>
<td>33.8 (3.6)</td>
</tr>
<tr>
<td>SCOPE-UE</td>
<td>91.7 (1.5)</td>
</tr>
<tr>
<td>SYMMETRY</td>
<td>29.1 (4)</td>
</tr>
<tr>
<td>VALID CONTROLS 1</td>
<td>91.7 (2.7)</td>
</tr>
<tr>
<td>VALID CONTROLS 2</td>
<td>89.8 (2.1)</td>
</tr>
</tbody>
</table>

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

**Evidence for $\exists$-reading**  Statistical analyses on data from the FALSE and WEAK conditions revealed a significant effect of CONDITION ($\chi^2(1) = 67.7, p < .001$). There is thus evidence for the $\exists$-reading with all.

**Evidence for $\forall$-reading**  Statistical analyses on data from the STRONG and WEAK conditions revealed a significant effect of CONDITION ($\chi^2(1) = 49.2, p < .001$). There is thus evidence for the $\forall$-reading with all.

**Between-participants variation**  There is also noticeable variation in terms of how the participants judged the WEAK condition, as evidenced by the distribution of the
participants’ means in Figure 6. This shows that speakers differ in the extent to which they access the $\exists$-reading and the $\forall$-reading.

![Graph showing distribution of participants' mean responses in Weak condition]

Figure 6: Exp. 3: Distribution of participants’ mean responses in Weak

### A.6 Subjective monotonicity and the amount of $\forall$-reading

The first question we wanted to address in Experiment 3 is whether each participant’s subjective perception of monotonicity of the quantifier *all* (to be computed based on their responses in Inference judgment task) predicts the rate at which that participant accesses the $\forall$-reading of donkey anaphora (to be computed based on their responses in Truth value judgment task).

According to Kanazawa’s generalization, sentences headed by $\uparrow \text{MON} \uparrow$ or $\downarrow \text{MON} \downarrow$ have the $\exists$-reading as default, sentences headed by $\uparrow \text{MON} \downarrow$ or $\downarrow \text{MON} \uparrow$ have the $\forall$-reading as default reading; and quantifiers with other monotonicity profiles (non-monotonic in restrictor or scope) should not have a default. This last sentence is not really correct. K’s generalization is simply silent about non-monotonic cases, and he in fact explicitly claims that exactly $n$ is $\exists$-default. When it comes to inferential properties, it might not be too farfetched to expect non-monotonic environments to be more similar to downward-entailing environments and to upward-entailing environments than these two are to each other. This is because downward-entailing environments and upward-entailing environments license opposing inferences (superset to subset in the former case, and subset to superset in the latter case), while non-monotonic environments share with downward-entailing environments the property of not licensing subset to superset inferences, and with upward-entailing environments the property of not licensing superset to subset inferences. This further means that another way to state Kanazawa’s generalization is as in (33):

$$\text{(33) Restating Kanazawa’s generalization:} \text{ The larger the difference in monotonicity properties between the restrictor and the scope of a quantifier, the more easily that quantifier receives the } \forall \text{-reading of donkey anaphora.}$$
According to this generalization, if the preservation of 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 \( \forall \)-reading of donkey anaphora in sentences headed by all.

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

In order to correct for different uses of the response scale among the participants, we normalized for each participant \( p \) their responses \( r \) on items in Weak as in (34), where \( \text{mean}_p(\text{Strong}) \) and \( \text{mean}_p(\text{False}) \) 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.

(34) \[
\frac{r - \text{mean}_p(\text{False})}{\text{mean}_p(\text{Strong}) - \text{mean}_p(\text{False})}
\]

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, 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. The normalization procedure in the inference judgment task was completely parallel to that in the Truth value judgment task. Third, we calculated the mean of directional responses in each of the four monotonicity-related conditions, and computed subjective perception of monotonicity of the scope for each participant as in (35a), and the subjective perception of monotonicity of the restrictor as in (35b), with \( \text{mean}_p(\text{dir-Scope-DE}) \) the mean of 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 (35a), and the monotonicity of the restrictor, calculated as in (35b).

(35) 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 see whether or not the monotonicity index predicts the incidence of the \( \forall \)-reading of donkey anaphora, the data was subsetted to items in \textsc{Weak}. A linear mixed effect model was fitted on the responses normalized to the items in \textsc{Weak}, with \textsc{monotonicity index} as a fixed effect, and by-participant intercepts as a mixed effect. A comparison of this model with a reduced model without \textsc{monotonicity index} as a fixed effect revealed no significant effect (\( \chi^2(1) = 0.005, p = .94 \)). In other words, there is no evidence that subjective perception of monotonicity predicts the amount of \( \forall \)-readings of donkey anaphora.

For completeness, we computed the by-participant average of normalized responses in \textsc{Weak} in the truth value judgment task, and plotted them as a function of the monotonicity indices in Figure 7a.

![Graphs showing the relationship between monotonicity and \( \forall \)-reading, and symmetry and \( \forall \)-reading.](image)

(a) Monotonicity and \( \forall \)-reading  \hspace{1cm} (b) Symmetry and \( \forall \)-reading

Figure 7: Participants’ averages of normalized responses in \textsc{Weak} in the truth value judgment task as a function of their monotonicity indices in (a) and of their symmetry indices in (b). Note that the higher the average in \textsc{Weak}, the lower incidence of the \( \forall \)-reading, because the \( \forall \)-reading is not verified in \textsc{Weak}.

A.7 Subjective symmetry and the amount of \( \forall \)-reading

The second question we wanted to address in Experiment 3 is whether or not each participant’s subjective perception of symmetry of the quantifier \textit{all} (to be computed based on their responses in the inference judgment task) predicts the rate at which that participant accesses the \( \forall \)-reading of donkey anaphora.

We have observed that asymmetric quantifiers appear to allow more readily for the \( \forall \)-readings of donkey anaphora. From this 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 $\exists$-readings of donkey anaphora it should give rise to.

We have already established how to evaluate the extent to which each participant accesses the $\forall$-reading in Section A.6. 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.6). Each participant’s *symmetry index* is calculated as this participant’s mean of the normalized responses in SYMMETRY.

To see whether or not the symmetry indices predict the robustness of the $\forall$ 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 by-participant intercepts as a mixed effect. 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 $\forall$-readings of donkey anaphora.

For completeness, we computed each participant’s average of normalized responses in WEAK in the truth value judgment task, and plotted them as a function of their symmetry indices in 7b.

### A.8 Discussion of Experiment 3

In Experiment 3, we investigated whether or not subjective perception of monotonicity and of symmetry of the quantifier *all* predicts the interpretation of donkey anaphora with this quantifier. We found considerable variation in the amount of $\forall$-readings among different participants (cf. Figure 4a), but this variation does not correlate with their subjective monotonicity and symmetry of the quantifier *all*. This fact is a challenge for Kanazawa’s (1994) theory according to which the interpretation of donkey anaphora in the scope of a quantifier is selected so that it preserves the inferential patterns based on monotonicity and/or symmetry of the quantifier in question.

### References


