Cognition 273 (2026) 106523

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Full length article

Adult learning of a novel quantifier tracks semantic universals

Sonia Ramotowska [b,a,*], Leendert van Maanen [c], Jakub Szymanik [d]

[a] Institut Jean-Nicod, Département d'Études Cognitives, École Normale Supérieure - PSL, EHESS, CNRS, Paris, France
[b] Institute for Logic, Language and Computation, University of Amsterdam, The Netherlands
[c] Department of Experimental Psychology & Helmholtz Institute, Utrecht University, The Netherlands
[d] Center for Mind/Brain Sciences and Department of Information Engineering and Computer Science, University of Trento, Italy

ARTICLE INFO

Dataset link: https://osf.io/gv25w

Keywords:
Quantifiers
Semantics
Lexical universals
Learnability
Monotonicity
Convexity
Conservativity
Quantity

ABSTRACT

Natural languages share common properties called universals. In the domain of quantification, three semantic universals were discovered: monotonicity (convexity), quantity, and conservativity. Researchers have been trying to explain the origin of semantic universals for decades. In this study, we tested one of the proposed explanations, the learnability hypothesis. According to this hypothesis, quantifiers that satisfy universals are easier to learn and, therefore, more likely to be lexicalized in natural language. We tested the learnability hypothesis in a large-scale, online, between-subjects design experiment, in which participants learned a new quantifier, *gleeb*. *Gleeb* corresponded to one of the following quantifiers: monotone, convex, and quantitative *at least 3* and *at most 2*, non-monotone and convex *between 3 and 6*, non-convex *even number*, non-quantitative *the first 3*, *the last 3*, conservative *not all*, and non-conservative *not only*. We found that all universals affected the speed of acquisition: participants learned conservative quantifiers faster than non-conservative quantifiers, quantitative quantifiers faster than non-quantitative quantifiers, and upward monotone quantifiers faster than non-monotone quantifiers (however, the lack of convexity did not affect the rate of learning more). In conclusion, our results provide evidence for the learnability hypothesis.

1. Introduction

Natural languages share common properties. For example, at the phonological level, it turns out that "all spoken languages have consonants and vowels" (Hyman, 2008). At the lexicon level, it is true that "every human language has proper names" (Universal 1321 Hockett, 1963). In the number system domain, Greenberg (1978) proposed the universal "in every numerical system, some numbers are expressed by basic terms" (Universal 528). Explaining the origins of universal properties is one of the main goals of linguistic theory and the cognitive science of language. In this paper, we focus on what has historically been considered perhaps the most prominent universals at the semantic level of linguistic analysis.

Quantifiers[1] are logical expressions that denote the relationship between sets (or quantities), for example: *most*, *at least 6*, *all*, *at most 10*, *between 11 and 19*, etc. Some quantifiers are lexicalized in natural languages as **determiners**. For example, in English, *most*, *all*, *some*, *few*, and *none* are lexicalized, but the quantifier that expresses a quantity *between 2 and 10* is not lexicalized. In contrast, this quantifier is lexicalized as *kilka* in Polish (Szymanik & Kieraś, 2022). In combination with common nouns, determiners create noun phrases (restrictors), e.g., *most dogs*, *some cats*, *few students*. When combined with a verb phrase (scope), they constitute quantified sentences, e.g., *most dogs chase cats*, *some cats like fish*, *few students are brilliant*.

Cross-linguistic research has revealed that those quantifiers that are lexicalized in the world's languages tend to obey certain semantic universal constraints (Barwise & Cooper, 1981; Keenan & Stavi, 1986; Peters & Westerståhl, 2008):

Convexity. Lexical quantifiers denote contiguous quantities without gaps. A non-convex quantifier like *at most 2 or at least 11* has a gapped denotation (it is true for quantities smaller than three and quantities larger than ten, but false for numbers between three and ten, hence has a gap). Such expressions are not observed as lexicalized determiners in natural languages. In contrast, its negation, the above-mentioned Polish

* Correspondence to: Institut Jean Nicod, 29 Rue d'Ulm, 75005 Paris, France
E-mail address: sonia.ramotowska@ens.psl.eu (S. Ramotowska).

[1] In this paper, we focus only on quantifiers of type ⟨1, 1⟩, see Peters and Westerståhl (2008), Szymanik (2016).

https://doi.org/10.1016/j.cognition.2026.106523
Received 4 July 2025; Received in revised form 5 March 2026; Accepted 7 March 2026
Available online 15 March 2026

determiner *kilka*, denotes a contiguous intermediate range. A stronger version of convexity[2] constrained is **monotonicity**.[3] Upward monotonicity means that enlarging the set of quantified objects preserves the quantifier's truth value, while downward monotonicity means the opposite. For example, *at least 3* is true if the set of objects is increased to four, five, etc. Most lexical quantifiers are either upward or downward monotone (e.g., *all, some, no*), and virtually all are convex. Barwise and Cooper (1981) formulated this observation as a semantic universal:

All lexicalized determiners express monotone quantifiers or conjunctions thereof.

Quantity. Quantity[4] (van Benthem, 1986) is traditionally a part of the mathematical definition of generalized quantifiers (Mostowski, 1957). The truth value of a lexical quantifier depends only on the number of items in certain sets, not on other properties like their identity or order. In other words, natural language determiners express purely numerical relations. For example, the truth value of a complex expression such as *the first 3*, in addition to quantity, depend on the sequential position of quantified elements, while the truth value of determiners like *three, some*, or *most* depend solely on counts (e.g., "three cats" is true based on the number of cats and does not require those cats to be the first three in any order). The non-quantitative quantifiers are not lexicalized as single determiners in natural languages. This phenomenon has a status of the quantity universal (Peters & Westerståhl, 2008):

All lexicalized determiners are quantitative.

Conservativity. Conservativity states that a determiner meaning only evaluates the relationship between a set A (the restrictor) and its intersection with another set B (the scope), ignoring elements outside A.[5] For example, "All A are B" is true if and only if "All A are $(A \cap B)$" is true – any B that is not an A is irrelevant to the truth of the statement. This property holds for all natural-language determiners: e.g., "All cats are pets" only concerns cats, and whether non-cats are pets has no bearing on its truth. A non-conservative quantifier violates this restriction, requiring consideration of individuals outside the restrictor set. For instance, *not only* (as in "Not only cats are pets") is true precisely when some non-A (non-cats) are included in B (are pets) – one must look beyond set A to evaluate it. Such non-conservative determiner meanings (e.g., *not only*, or an artificial determiner that is true when Bs are a subset of A) are unattested as simple lexical items in any language. Conservativity, perhaps the most famous among quantifier universals (Barwise & Cooper, 1981; Keenan & Stavi, 1986; see Peters & Westerståhl, 2008 for discussion),[6] states:

All determiners are conservative.

Goals of the paper. While these generalizations accurately describe cross-linguistic patterns, their origin remains a matter of debate. The key theoretical question is why these particular semantic universals exist: what forces or constraints favor convexity, quantity, and conservativity in quantifier systems? Researchers have proposed multiple explanations, invoking pressures such as communication efficiency (Steinert-Threlkeld, 2021), simplicity bias (van de Pol et al., 2023; Steinert-Threlkeld, 2020), grammatical interface (Romoli, 2015), and learnability (Carcassi et al., 2021; Steinert-Threlkeld & Szymanik, 2019). In this paper, we investigate the hypothesis that these universals arise, in part, from the cognitive constraints of human learning and concept acquisition. According to this **learnability hypothesis**, quantifier meanings that obey the universal constraints are inherently easier for humans to learn, and therefore more likely to be lexicalized as words across languages. This idea builds on the intuitive notion that *easier to learn meanings tend to be lexicalized*. Notably, various theoretical and modeling studies have postulated or exemplified this hypothesis in the context of quantifiers. For example, certain computational models show faster learning of monotonic or quantitative quantifiers, and iterated learning simulations have demonstrated the emergence of monotone quantifiers in artificial languages over generations of learners (Carcassi et al., 2021; Steinert-Threlkeld & Szymanik, 2019, 2020). These computational works lent plausibility to the learnability account but did not constitute a direct empirical test with human learners. Here, we present the first comprehensive large-scale experimental test of the learnability hypothesis in adult humans for all quantifier universals. Our study contributes to the empirical validation of the learnability hypothesis. In what follows, we present the partial evidence for the learnability hypothesis and novel experimental data supporting this hypothesis. In this study, we do not specify one psycholinguistic mechanism of learning, as was done in the experiments with computational models (Steinert-Threlkeld & Szymanik, 2019). However, we briefly analyze some candidates for these mechanisms. Finally, we discuss how our results fit on the topographic map of alternative explanations.

1.1. Evidence for the learnability hypothesis

1.1.1. Quantity

To our knowledge, the effect of quantity on learnability was not tested in human participants, but the artificial learning model learned the quantitative quantifier *at least 3* faster than the two non-quantitative quantifiers *the first 3* and *the last 3* (Steinert-Threlkeld & Szymanik, 2019). More experimental and computational evidence for the learnability hypothesis comes from studies investigating its role in the formation of monotonicity (Chemla, Buccola, et al., 2019) and conservativity (Hunter & Lidz, 2013) universals.

1.1.2. Monotonicity and convexity

Partial empirical evidence for a role of monotonicity and convexity universals in the learnability of quantifiers comes from experiments on adult humans (Chemla, Buccola, et al., 2019). In the former experiment, participants saw a display with five circles in different colors and were asked to assess whether the display was consistent with one of the three types of rules: monotone (e.g., "There are 3, 4, or 5 red circles."), connected (e.g., "There are 2, 3, or 4 red circles."), or non-connected (e.g., "There are 1, 2, or 4 red circles."). Participants had to infer the rule based on feedback in each trial. Chemla, Buccola, et al. (2019) showed that participants learned the non-convex rule more slowly than the monotone rule.

Moreover, Steinert-Threlkeld and Szymanik (2019) showed the monotonicity role in learnability by using a long short-term memory recurrent neural network model for learning. The network learned faster both upward (*at least 4*) and downward (*at most 3*) monotone quantifiers than non-monotone quantifiers (*at least 6 or at most 2*).

Finally, additional evidence for the role of the monotonicity universal comes from language-evolution experiments. Carcassi et al. (2021)

[2] Convexity is also referred to as connectedness (Chemla, Buccola, et al., 2019). Peters and Westerståhl (2008) use the term *continuous quantifiers* for convex quantifiers, p. 168. Formally, convexity can be defined as in Chemla, Buccola, et al. (2019) (Q is a quantifier, A and B are sets, and M is a domain):

(1) The Q is convex iff if $\langle M, A, B' \rangle \in Q$, $\langle M, A, B'' \rangle \in Q$ and $B' \subseteq B \subseteq B''$ then $\langle M, A, B \rangle \in Q$.

[3] We call a quantifier monotone if and only if it is either upward monotone or downward monotone. Formally (Q is quantifiers, A and B are sets, and M is a domain):

(2) The Q is upward monotone iff if $\langle M, A, B \rangle \in Q$ and $B \subseteq B'$, then $\langle M, A, B' \rangle \in Q$.

(3) The Q is downward monotone iff if $\langle M, A, B \rangle \in Q$ and $B' \subseteq B$, then $\langle M, A, B' \rangle \in Q$.

[4] Formally, quantity (Peters & Westerståhl, 2008) is defined as isomorphism invariance (Q is a quantifier, A and B are sets, and M is a domain):

(4) The Q is isomorphism-invariant iff if $\langle M, A, B \rangle \cong \langle M', A', B' \rangle$, then $\langle M, A, B \rangle \in Q$ if and only if $\langle M', A', B' \rangle \in Q$.

[5] Formally, conservativity is defined as (Q is a quantifier, A and B are sets):

(5) The Q is conservative iff: $\langle M, A, B \rangle \in Q$ iff $\langle M, A, A \cap B \rangle \in Q$.

[6] Barwise and Cooper (1981) did not use the term "conservativity". They formulated the determiner universal (Universal 3) on p. 179. The term "conservativity universal" was first used by Keenan and Stavi (1986).

showed the emergence of monotone quantifiers in their iterative learning experiment with neural network agents. Moreover, the monotonicity universal plays a vital role in other language domains; for example, scalar adjectives (Carcassi et al., 2019). In the iterated language experiment, Carcassi et al. (2019) showed that learnability is one of the pressures for monotonicity to evolve. They fitted three computational models within the Rational Speech Act framework (Frank & Goodman, 2012) and showed that the model with combined learnability pressure and agents' pragmatic skills led to the evolution of monotone adjectives.

1.1.3. Conservativity

The role of learnability in explaining conservativity as a semantic universal is the most controversial. Some studies have shown that conservativity facilitates learning in children (Hunter & Lidz, 2013), while others have not replicated this finding (Spenader & de Villiers, 2019). Hunter and Lidz (2013) showed that children learned the artificial quantifier *gleeb* faster when its meaning corresponded to the conservative quantifier *not all* than to the non-conservative quantifier *not only*. Spenader and de Villiers (2019) failed to replicate Hunter and Lidz's 2013 finding in both adults and children. Neither group was successful in learning the new quantifier. Importantly, they were equally unsuccessful, regardless of whether the quantifier satisfied the conservativity universal. Moreover, Spenader and de Villiers (2019) applied a different experimental paradigm, a situation verification with correction, to further test the effect of conservativity on learnability. They found that children learned the conservative quantifier *all* and the non-conservative quantifier *only* equally well, but did not learn the conservative quantifier *not all* and the non-conservative quantifier *not only*. All four experiments failed to show the difference between conservative and non-conservative quantifiers. It is worth mentioning that Hunter and Lidz's 2013 experiment consisted of training and test blocks, both very short, with only five trials. Spenader and de Villiers's 2019 novel paradigm, in turn, consisted of 10 training and 10 testing trials. The experiment was therefore also fairly short, and, as the authors admitted, the lack of difference between the quantifiers *not all* and *not only* could be due to insufficient training trials. Hence, participants might not have had enough time to become familiar with the experiment and learn the quantifier simultaneously. This means that the test data may have contained substantial noise unrelated to the tested universal (for example, because participants could have made many response errors). As a consequence, these studies show inconsistent results. Moreover, both experiments were highly underpowered. Hunter and Lidz (2013) recruited 10 participants per quantifier, whereas in a replication experiment, Spenader and de Villiers (2019) included only 9 participants per quantifier.

In the series of four experiments, Knowlton et al. (2022) showed that adult participants learned conservative determiners and verbs better than non-conservative ones, even after being explicitly taught the meaning of a non-conservative novel word. Moreover, they showed that participants performed below chance for non-conservative determiners, which the authors interpreted as evidence for the unlearnability of novel non-conservative meanings. The study provided strong evidence for a learnability hypothesis as a source of conservativity.

Steinert-Threlkeld and Szymanik (2019), in turn, suggested that conservativity may be explained independently of learning simplicity. In their experiments, the neural networks learned conservative and non-conservative quantifiers: *not all* and *not only*, as well as the pair *most* and an artificial non-conservative quantifier *M* (meaning of *M* was $|A| > |B|$). Both types of quantifiers were learned at the same rate. However, the authors admit that this does not constitute proof of the equal learning complexity of the two quantifiers. In fact, the encoding used by the authors enforced the result.[7]

1.2. What is learning?

The learnability hypothesis is fundamentally agnostic regarding the specific cognitive nature of the learning process. Evidence for this hypothesis spans diverse paradigms and agents, ranging from artificial neural networks with no prior linguistic knowledge (Steinert-Threlkeld & Szymanik, 2019) to children acquiring their first language (Hunter & Lidz, 2013; Spenader & de Villiers, 2019). In some cases, learning is operationalized as implicit rule induction (Chemla, Buccola, et al., 2019), while in others, it involves the lexicalization of a new word-concept (Knowlton et al., 2022). Consequently, *learning* must be defined relative to the specific constraints and prior knowledge of the agent in question.

In the case of human adults, next to implicit learning, a plausible learning mechanism is a search through a structured hypothesis space of possible meanings. Adult participants already possess a sophisticated linguistic repertoire biased toward semantic universals. When tasked with learning a novel word like *gleeb*, these agents likely engage in a process of *mapping*: testing the input against existing mental concepts or logical combinations of those concepts. Under this framework, the learnability hypothesis predicts that the efficiency of this search-and-mapping process is modulated by universal properties. Concepts satisfying universals, such as monotonicity or conservativity, are hypothesized to be more accessible or less complex in the human mental lexicon. From this perspective, mapping is not an alternative to learning, but rather the specific mechanism by which adult learners converge on the meaning of a new lexical item.

1.3. Current experiment

The goal of this study was to test whether the typological generalization, as defined by the three quantifier universals, is consistent with the learnability hypothesis. If learnability explains the origin of semantic universals in the domain of quantification, we should observe that participants learn quantifiers that satisfy universal properties more easily than those that do not. In contrast, if the semantic universals of interest are not related to learnability, we should not observe a difference in learning.

We adopted an artificial-word learning experimental paradigm (e.g., Maldonado & Culbertson, 2022; Maldonado et al., 2022). In our experiment, participants learned the meaning of a new word, *gleeb*, by experiencing the conditions that satisfy and do not satisfy *gleeb*. *Gleeb* was presented with the partitive construction "of the" to ensure the determiner interpretation (see Knowlton et al., 2022 for discussion). We included two monotone quantifiers: *at least 3* (upward) and *at most 2* (downward), as well as non-monotone quantifiers *between 3 and 6* and *an even number*. We tested the convex quantifier *between 3 and 6* versus the non-convex quantifier, *even number*. For quantity, we compared the quantitative quantifier *at least 3* with the non-quantitative quantifiers *the first 3* and *the last 3*. For conservativity, we chose *not all* (conservative) versus *not only* (non-conservative). All of the chosen quantifiers are complex quantifier expressions that are not lexicalized in English. We chose non-lexicalized quantifiers to equate the learning task between universals satisfying and not satisfying quantifiers. Because all lexicalized quantifiers satisfy the semantic universals, the learning task would naturally be more difficult for quantifiers that do not satisfy the universals. By using the complex quantifiers, we aimed to prevent participants from mapping the meaning of a new word to an existing *single* word in their language. Instead, what we tested in our experiment was rather the ability to conceptualize a complex quantifier as a single word.

While simulation studies can involve many quantifiers with different degrees of semantic universals,[8] experimental studies are limited

[7] See Steinert-Threlkeld and Szymanik (2019) for a technical explanation of this result.

[8] Carcassi et al. (2021) have introduced an information-theoretic way to turn the binary semantic properties into graded measures. The general recipe

to using the minimal pairs methodology. The idea is to choose pairs of quantifiers that differ in universal properties but are otherwise comparable and test them against each other. The choice of quantifiers for this experiment was motivated by previous studies (Chemla, Buccola, et al., 2019; Hunter & Lidz, 2013; Spenader & de Villiers, 2019; Steinert-Threlkeld & Szymanik, 2019), to make our results comparable with those previously reported. However, in contrast to previous studies (e.g., Ramotowska, 2022; Steinert-Threlkeld & Szymanik, 2019), we used only quantifiers attested in corpus data (see reported frequencies in Appendix A.2). This was intended to equate the differences between universal satisfying and non-satisfying quantifiers as much as possible. We chose numerical monotone, convex, and non-convex quantifiers to test quantifiers with different degrees of monotonicity (Carcassi et al., 2021). In this way, we tested the same contrast monotone vs. non-monotone/convex vs. non-convex as Chemla, Buccola, et al. (2019). We chose the same non-quantitative quantifiers as Steinert-Threlkeld and Szymanik (2019). The non-quantitative quantifiers in their study referred to numerical information and the order of presentation. For example, the sentence "*The first 3* of the triangles are red" requires a presentation of *at least 3* red triangles, as well as ordering. There must be three red triangles at the beginning to satisfy the sentence. Therefore, we contrasted the non-quantitative quantifiers with the monotone quantifier *at least 3*.[9] We chose the same conservative vs. non-conservative pair of quantifiers as Hunter and Lidz (2013), Spenader and de Villiers (2019), and Steinert-Threlkeld and Szymanik (2019). We selected this pair because it is comparable in terms of complexity (van de Pol et al., 2023) and because both quantifiers are present in natural language.[10]

In comparison to previous experiments, we introduced several methodological advancements. We adopted the same experimental paradigm to test differences between quantifiers that do and do not satisfy the universals, and to compare different universals. This uniform methodology allowed us to systematically investigate the learnability effect across universals and factor out the effect of the experimental paradigm. Additionally, we included the quantity universal, which was not previously tested in humans. Studies with a very small sample size, such as Hunter and Lidz (2013), Spenader and de Villiers (2019), might lead to inconsistent results (cf. Aarts et al., 2015). One advantage of online experiments is the possibility to collect a large amount of data in a short time (Kochari, 2019). We used this opportunity to collect substantially larger sample sizes per quantifier and obtain more reliable results.

2. Methods

2.1. Participants

395 participants took part in our experiment, which was implemented in PsychoPy and PsychoJS version 2023.2.3 and stored on the Pavlovia platform (https://pavlovia.org/). Participants were recruited via the Prolific platform (https://www.prolific.co/), where they were provided with the link to the experiment. As an attention check, we used an easy secondary task, in which participants had to indicate a correct geometric shape. We excluded 23 participants who made at least one mistake in this task (see Appendix A.3 for control analysis without exclusion). The final sample consisted of 372 participants: *at least 3* (N = 49), *at most 2* (N = 44), *between 3 and 6* (N = 46), *even number* (N = 46), *the first 3* (N = 50), *the last 3* (N = 47), *not all* (N = 44), *not only* (N = 46). In addition, 5 participants did not provide typed responses at the end of the experiment. The mean age in the final sample was 41 years (range 18 – 78), 120 participants were male, 251 female, and one was non-binary. All participants were self-reported native speakers of English and self-reported non-colorblind. The Research Ethics Committee of the University of Trento accepted the experimental procedure. Participants were paid a 9£hourly rate.

2.2. Materials and design

The experiment had a between-subjects design, meaning each participant learned only one quantifier. It consisted of the secondary and learning tasks. The learning task was split into two parts: training, in which participants learned the meaning of the quantifier with feedback, and testing, in which they did not receive feedback about their responses. The stimuli consisted of a display containing a sentence "Gleeb of the SHAPES are COLOR", where SHAPES corresponded to triangles, squares, or circles and COLORS to red, green, yellow, or blue and between one and eight geometric objects positioned in the row below the sentence (see Fig. 1 for example stimuli). The objects were displayed in a row because their order was necessary for non-quantitative quantifiers. In contrast to relatively simple previous designs, our experiment had a high degree of stimulus variation: the number of geometric shapes, the types of shapes depicted, and their colors. The variation was intended to ensure that participants could not simply memorize the stimulus–response mapping, and they actually had to learn the meaning of the new word. In addition, the variation in the quantifier restrictor and scope was intended to draw participants' attention to the sentence. The stimuli for monotonicity and quantity minimality differed from those for conservativity, in that the former included only one type of shape, whereas the latter included two (necessary for testing non-conservative quantifiers). We used only one shape type for monotonicity and quantity to balance variability across stimuli and their visual clarity. In the following subsections, we describe each quantifier's stimuli in more detail.

2.2.1. Learning task stimuli: monotonicity, convexity, and quantity

For quantifiers *at least 3*, *at most 2*, *between 3 and 6*, *even number*, *the first 3*, and *the last 3*, we used one type of figure in each trial. We randomly generated a list of 100,000 sequences with between zero and eight targets and between zero and seven fillers and their order permutations (e.g., target – filler – target vs. filler – target – target). The repetitions of the same sequences were removed. Then we assigned the correct response to each sequence based on the quantifier's meaning. Because the number of possible trials per target numerosity was unequal (e.g., only one for eight targets), we had to apply undersampling and oversampling procedures. For each block, we randomly drew six trials in which the quantifier was true and six in which it was false, ensuring that all target numerosities (0–8) were represented. In this way, we ensured that participants saw each numerosity at least 7 times during training (across 7 blocks). In addition, we assessed how distinctive a given sequence is for a given quantifier. For example, the quantifier *at least 3* is true in many instances in which *most* or *some* are also true. Thus, for each sequence, we computed its probability of being true given the most frequently used quantifiers: *all*, *most*, *some*, *none*, and *at least half*, and we selected sequences with various ranges of probabilities between 0.2 and 0.8 (see Appendix A.1). Finally, the shapes, target colors (the color mentioned in the sentence), and filler

is to treat the pairs of models relevant to the universal (e.g., a model and its supermodels for upward monotonicity, or a model and its restriction for conservativity) as random variables, and then to quantify the dependence between their truth-values using normalized mutual information, yielding maximal dependence for fully constrained quantifiers and graded degrees for non-constrained ones.

[9] Note that all monotone (*at least 3* and *at most 2*) and non-quantitative (*the first 3* and *the last 3*) quantifiers are conservative, and that non-quantitative quantifiers are monotone. For example, from the sentence "*The first 3* triangles are light red", it follows that "*The first 3* triangles are red". Therefore, *the first 3* is monotone increasing.

[10] In addition, Steinert-Threlkeld and Szymanik (2019) tested another pair of conservative and non-conservative quantifiers, the pair *most* and *M*, which we did not include in our study, because *most* is already lexicalized in English, while *M* is an artificial quantifier. Nonetheless, Steinert-Threlkeld and Szymanik (2019) showed that the results for the *not all* and *not only* pair are qualitatively identical to the results for *most* and *M*. See the discussion of the limitations of that experimental comparison above.

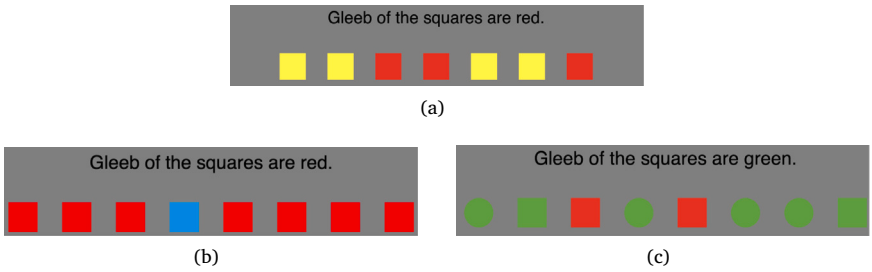

Fig. 1. Example stimuli used in the experiment. Figs. 1(a) and 1(b) show example stimuli for monotone, convex, non-convex, and non-quantitative quantifiers, contrasting the stimulus in which the order of squares was random with the stimulus for non-quantitative quantifiers in which the order of squares was important. 1(a) shows seven squares (three red) in a random order. For this stimulus, participants should have provided a "true" response for quantifiers: *at least 3* and *between 3 and 6*, and a response "false" for *at most 2* and *even number*, *the first 3*, and *at least 3*. 1(b) shows eight squares (seven red), the first three red. For this stimulus, participants should have provided a "true" response for quantifiers: *the first 3*, *the last 3*, and *at least 3*, and a response "false" for *at most 2* and *even number*. 1(c) shows an example stimulus for conservative vs. non-conservative quantifiers, eight figures (four squares, four circles). For this stimulus, participants should have responded "true" to the quantifiers *not all* and *not only*. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

colors (a different color from the target color) were randomly drawn for the final list of trials. The objects were centered on the screen. Each participant was trained on the same set of trials, and the order of blocks was randomized.

A separate set of 12 trials was generated using the same procedure for the testing block.

2.2.2. Learning task stimuli: conservativity

For conservativity, we used two types of figures, with a maximum of four of each type. To avoid presupposition violation for *not only* and implicature violation for *not all*,[11] one figure of the target type in the target color was always displayed. The training and testing trial generation procedure was the same as for other quantifiers, ensuring consistency.

2.3. Procedure

At the beginning of the experiment, participants provided basic demographic information (gender and age) and gave informed consent to participate. The experiment consisted of three parts: secondary task, training, and testing. All participants completed the experiment in this order.

In the secondary task, participants were presented with two squares in different colors, one on the right side and one on the left side of the screen. They had to indicate, "Which square is COLOR?', where COLOR was green, yellow, blue, or red, by pressing the R key for right or the L key for left. We tested all six color combinations. The target color appeared three times on the left and three times on the right.

The training part of the main task consisted of 84 trials, 7 blocks of 12 trials. We included extensive training to reduce data noise. We scaled the length of the experiment based on findings by Chemla, Buccola, et al. (2019). We decided that 84 trials should be enough for participants to learn even more difficult quantifiers, while keeping the experiment short enough to avoid a high drop-out rate (Schnoebelen & Kuperman, 2010).

Each trial of the training part of the experiment consisted of two displays. On the first screen, participants saw a sentence, e.g., "Gleeb of the squares are red", and a depicted situation that the sentence described. The participants' task was to evaluate whether the sentence was true or false based on the situation. They had to press the T key if they thought the sentence was true or the F key if they thought it was false.

On the second screen, participants saw the same geometric objects (in the same colors and position on the screen) and feedback about their answers. They saw feedback "Correct!" displayed in green if their answer was correct and "Wrong!" displayed in red if their answer was incorrect. Correct-response feedback was displayed for 1 s, and incorrect-response feedback was displayed for 3 s to encourage participants to provide correct answers. After the feedback disappeared, the new trials started. There were an equal number of true and false trials in each block. We did not include breaks between blocks.

In the testing part of the experiment, the displays and stimuli were created in the same way as in the training, but participants did not receive feedback about their responses. In the instructions for the testing part, participants were told that the meaning of *gleeb* is the same in this part of the experiment as in the previous part. The testing consisted of one block (12 trials, 6 in which the correct response was true and 6 in which the correct response was false). After completing the experiment, the participants were asked about the meaning of *gleeb*.

2.4. Analysis of the typed responses

We asked two native English speakers with linguistic backgrounds to assess the typed responses from participants in the experiment. The consultants were told that in the experiment, participants learned the meaning of the new word *gleeb* used in sentences like "Gleeb of the triangles are red". Consultants had to decide, based on the typed responses, whether the subjects interpreted *gleeb* in the task as a determiner. They were given examples of common determiners such as definite and indefinite articles (*the, a*), quantifiers (*all, most, only*), numerical expressions (*three, more than k*), distributive determiners (*each, every*), or are part of complex noun phrases (*all triangles, three triangles*). Consultants used the following coding system, where 1 indicated that a response had a determiner interpretation of *gleeb*, 0 that the response did not have a determiner interpretation, and 999 if the response indicated that the participant did not know what *gleeb* meant. They were also provided with examples. To ensure the reliability of this analysis, we computed Cohen's Kappa statistic (using KAPPA2 function from *irr* R package Gamer et al., 2019).

3. Results

3.1. Main analyses

First, we counted how many participants performed above chance in the experiment's testing phase. This analysis was meant to test whether participants, based on the input data, could learn the meaning

[11] Previous studies have shown that the implicature *some* is robustly derived by many participants, see Ramotowska et al. (2024).

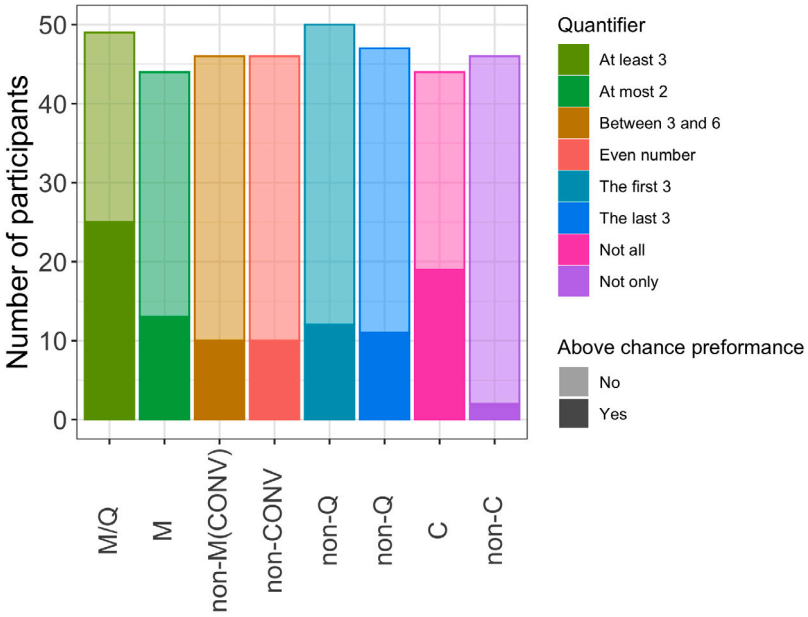

Fig. 2. The number of participants who performed above chance (darker color shade) or below chance (lighter color shade) in the test part of the experiment. The x-axis indicates whether the quantifier was included in analyses as universal satisfying (M/Q/C) or not (non-M/Q/C). The label non-M(CONV) indicates that the quantifier is convex but not monotone and the label non-CONV indicates that the quantifier is not convex. Labels indicate only the contrast included in the statistical analysis.

of all quantifiers included in this study. The above chance performance was determined by a significant one-sided binomial test ($p = 0.019$). Participants who performed above this threshold (83%; 10 out of 12 correct responses) were classified as above-chance-performing participants; see Fig. 2. *Not only* appeared to be the most difficult quantifier, as only 4% of participants performed above chance on this quantifier. In contrast, 43% of participants learned *not all*. 51% of participants learned monotone increasing quantifiers *at least 3*. The downward monotone quantifier seemed more difficult to learn; 30% of the participants crossed the threshold. 22% of the participants learned *between 3 and 6* and *even number* quantifiers. The non-quantitative quantifiers were learned at a similar rate (*the first 3* 24% and *the last 3* 23%). Together, the descriptive results suggest that participants could, in principle, learn the meaning of all quantifiers in this study; however, more participants were successful in learning quantifiers that satisfy universals than those that do not.

In the next step, we tested whether participants achieved higher accuracy in the testing part of the experiment for the quantifiers that satisfy universals than those that do not. The data from above-chance and below-chance performing participants were used. We ran mixed-effects logistic regression models for each universal (monotonicity/convexity, quantity, and conservativity), followed by pairwise comparisons with the Tukey correction of *p*-value for multiple comparisons. Dummy-coded quantifiers that satisfy or do not satisfy a given universal were included as the fixed effect, and the by-subject random intercept was included in all models. We defined the following pairwise contrasts. Model testing the monotonicity and convexity universal included *at least 3*, *at most 2*, *between 3 and 6*, and *even number*; model testing the quantity universal (*at least 3*, *the first 3*, *the last 3*); model testing the conservativity universal *not all* vs. *not only*. We used the *glmer* function from the *lmerTest* R package (Kuznetsova et al., 2017). Figs. 3–5 present the summary of the mean accuracy of participants in the testing part of the experiment and the accuracy of individual participants.

3.1.1. Monotonicity and convexity

For the monotonicity and convexity universals, the accuracy was the highest for *at least 3* (81%), lower for *at most 2* and *between 3 and 6* (both 67%), and the lowest for *even number* (58.2%). To investigate these differences statistically, we included *at least 3*, *at most 2*, *between 3 and 6*, and *even number* as predictors of the accuracy in the testing part of the experiment. The *post hoc* pairwise comparisons revealed that the accuracy for *at least 3* was significantly higher than the accuracy for *at most 2* ($\beta = 0.85; p < 0.0023$), *between 3 and 6* ($\beta = 0.90; p = 0.0008$) and *even number* ($\beta = 1.27; p < 0.0001$). In contrast, there was no significant difference between *at most 2* and *between 3 and 6* ($\beta = 0.05; p = 0.99$), and between *at most 2* and *even number* ($\beta = 0.42; p = 0.26$), or between *even number* and *between 3 and 6* ($\beta = 0.37; p = 0.36$).

3.1.2. Quantity

The accuracy was highest for *at least 3* (81%) and lower for *the first 3* (69.8%) and *the last 3* (70%). To test the accuracy differences between quantitative and non-quantitative quantifiers, we included the quantifiers *at least 3*, *the first 3*, and *the last 3* as predictors of response accuracy in the experiment's testing phase. The pairwise comparison revealed significant differences between quantifier pairs: *at least 3*, *the first 3* ($\beta = 0.66; p = 0.0008$), and *at least 3*, *the last 3* ($\beta = 0.65; p = 0.0014$), but not *the first 3*, *the last 3* ($\beta = -0.02; p = 0.99$).

3.1.3. Conservativity

The mean accuracy for *not only* was at chance level 52.5%, and for *not all* was higher at 66.6%. To test whether this difference in accuracy was significant, we included the quantifiers *not all* and *not only* as predictors of accuracy in the testing block of the experiment. The effect of the quantifier was significant ($\beta = 0.78; p = 0.002$).

To summarize, our results support the learnability hypothesis. Participants learned the monotone increasing quantifier better than non-monotone quantifiers, the quantitative quantifier better than the non-quantitative quantifiers, and the conservative quantifier better than the non-conservative quantifiers.

3.2. Analysis of typed responses

While our results seem to support the learnability hypothesis, we have not yet provided evidence that participants understood *gleeb* as a determiner. To answer this question, we analyzed the classifications of participants' typed responses made by two native English-speaking

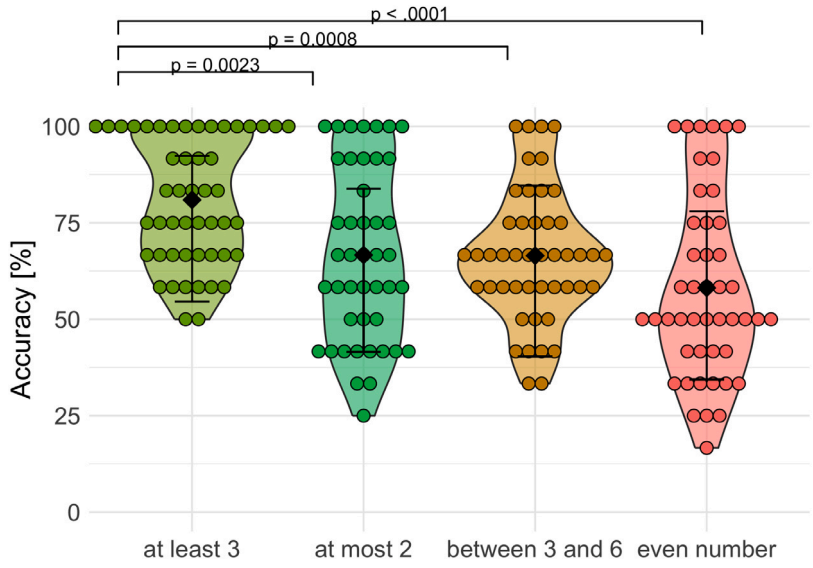

Fig. 3. Accuracy for *at least 3*, *at most 2*, *between 3 and 6*, and *even number*, mean with binomial 95% confidence interval.

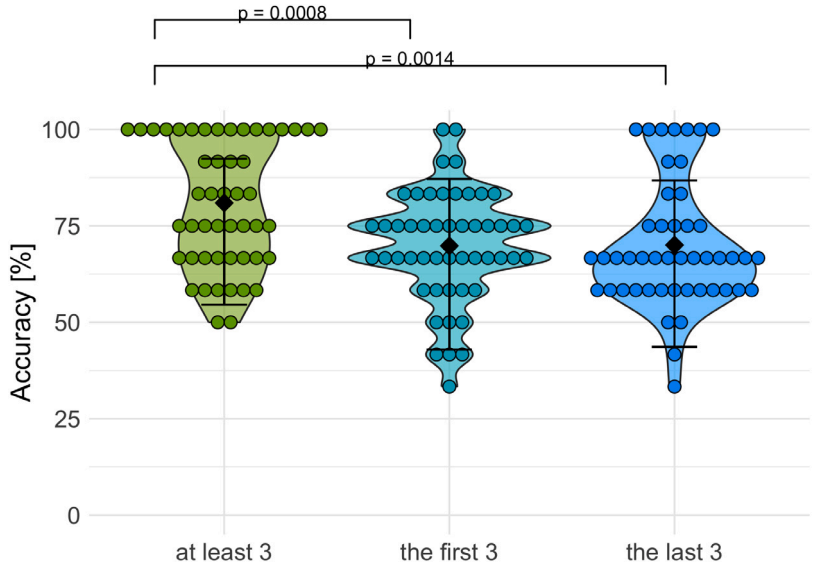

Fig. 4. Accuracy for *at least 3* and *the first/last 3*, mean with binomial 95% confidence interval.

consultants. 367 participants provided typed responses at the end of the experiment. 357 responses were classified by both consultants (one consultant was uncertain about the classification of 10 responses; these responses were coded as NA), yielding an unweighted Cohen's Kappa of 0.713, indicating substantial agreement. Of these 357 classifications, 210 responses were classified by both raters as determiner-like answers; 22 as a non-determiner, and 73 as "I don't know" answers.

We did not observe that the classifications among the cases for which the annotators agreed (see Table 1) were biased with respect to the universals. The non-determiner responses were evenly distributed across all quantifiers (1–5 responses). The highest number of "I don't know" responses was for *not only*, which conforms to this quantifier being the most difficult to learn.

The annotators disagreed about the classification of 52 typed responses. 17 of these cases were annotated by one consultant as determiner responses and by the second as "I don't know" responses. We checked that these responses contained a determiner and a phrase suggesting that participants were uncertain about the meaning of *gleeb* (e.g., "Still not sure - more than two?", "an even number? really not sure", or "No idea! I thought: half, none, some, more, less ..."). This suggests that for those cases, participants were seeking a determiner interpretation. 28 cases were classified by one annotator as a determiner and the other annotator as a non-determiner. For the remaining 7 cases, the disagreement was between classification as a non-determiner or "I don't know" responses.

Table 1
Distribution of 305 typed responses by quantifier for which both annotators agreed on classification: Det - determiner interpretation, non-Det - non-determiner interpretation.

Quantifier	Det	non-Det	I don't know
at least 3	37	2	3
at most 2	27	3	6
between 3 and 6	26	2	10
an even number	18	4	11
the first 3	27	4	9
the last 3	27	1	10
not all	29	1	8
not only	19	5	16

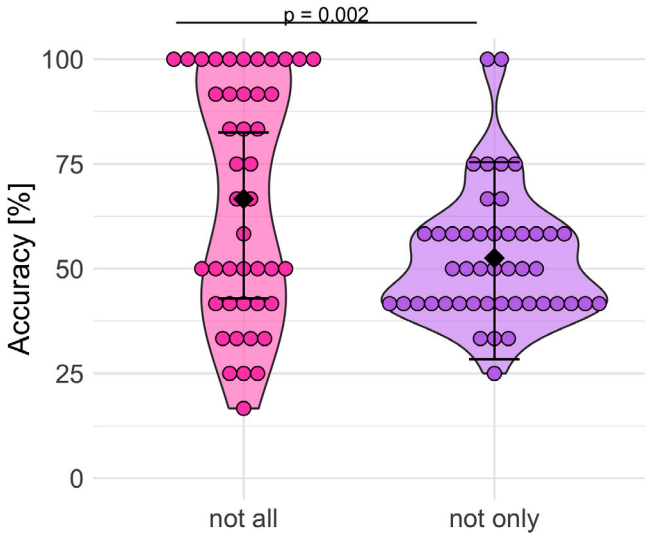

Fig. 5. Accuracy for *not all* and *not only*, mean with binomial 95% confidence interval.

We conclude that most participants in our experiment recognized the new word as a determiner. We believe that this result was reinforced by the use of the partitive construction.

3.3. Exploratory analyses

Even though in this study we do not investigate the specific learning mechanism, but we are rather focusing on the abstract notion of learning complexity, it is worth looking into one specific learning strategy in our setting: learning as the mapping between known concepts in participants' native language and the new word. Our main analyses showed that participants learned the concept that satisfies universals more successfully than the concept that does not. If we entertain the more specific mapping hypothesis, the two issues warrant discussion. Firstly, when searching for a concept expressed by *gleeb*, participants may have constrained their search to lexicalized concepts in their native language. If the meaning of these concepts was close to that of the "true" concept they were to learn, they could have achieved high accuracy in the learning task without learning the correct determiner. Moreover, because the lexicalized concepts also satisfy semantic universals, this approximation-learning strategy could have biased the results in favor of the tested hypothesis. Secondly, even though we identified mapping as a plausible learning mechanism in this study, participants could have correctly learn the meaning of a new word using different strategies, in particular, they could have learned *gleeb* implicitly (cf. discussion in Chemla, Buccola, et al., 2019).

To address these questions, we ran additional exploratory analyses that included typed responses as a control predictor of accuracy. While the data collected in this study are not diagnostic of specific learning strategies, we used a coarse-grained classification system and divided participants into explicit and implicit learner groups. The typed responses that at least one annotator classified as a determiner interpretation were diagnostic of an explicit learner group. All other participants were classified as implicit learners. Furthermore, we assessed the explicit learners' responses as correct or incorrect interpretations of *gleeb*. We noted that among these incorrect responses, quantifiers such as *most, the majority, some, all, none, half, or more than n*, were dominant, suggesting that participants were mapping the new word into lexicalized or frequent determiners in natural language. To test the effect of these strategies on learning accuracy, for each universal, we fit a mixed-effects logistic regression model with accuracy as the dependent variable and group classification (dummy-coded factor with the correct group as the reference level) and quantifier as predictors (a by-subject random intercept was also included). In short, although mapping seems to be broadly adopted by our successful subjects, the universal advantage persists when we control for learner types, meaning that all groups of learners find it harder to learn meaning-violating the universals. This suggests that the learnability hypothesis precedes or exists independently of the specific mapping mechanism.

3.3.1. Monotonicity

There were no significant interactions between the learner groups and quantifiers ($z_s = [-0.25, -1.77], p_s > 0.07$), meaning that the universal effect did not differ between learner groups. The model without interaction showed that implicit learners (51 participants, $\beta = -2.85, z = -9.39, p < 0.001$) and explicit incorrect learners (99 participants, $\beta = -2.81, z = -9.6, p < 0.001$) were less accurate than explicit correct learners (35 participants). This suggests that the mapping strategy was more effective than implicit learning, and that participants who identified the correct meaning of *gleeb* learned better than those who recognized the new word as a determiner but were unable to infer its true meaning. Pairwise comparisons revealed that *at least 3* was learned better than *at most 2* ($\beta = 0.66, z = 3.52, p = 0.002$) and *an even number* ($\beta = 0.89, z = 4.9, p < 0.001$) (*p*-values adjusted using the Tukey method for multiple comparisons). There was also a significant difference between *between 3 and 6* and *an even number* ($\beta = 0.51, z = 3.02, p = 0.01$). This suggests that the universal effect appears robust when controlling for a learner type. Fig. 6 visually demonstrates that the ease of acquisition is inherent to the quantifier property and not solely an artifact of successful mapping to English word-concepts.

3.3.2. Quantity

For quantity, the interactions between the learner groups and quantifiers were also not significant ($z_s = [-0.01, 1.45], p_s > 0.14$). As in the case of monotonicity, implicit learners (36 participants, $\beta = -2.11, z = -6.13, p < 0.001$) and explicit incorrect learners (88 participants, $\beta = -2.25, z = -6.82, p < 0.001$) were less accurate than explicit correct learners (22 participants). This effect is illustrated in Fig. 7. There was no effect of universal; however, a difference in learning between *at least 3* and *the last three* indicated a marginal effect ($\beta = -0.30, z = -1.83, p = 0.067$).

3.3.3. Conservativity

For conservativity, no significant interactions between the learner groups and quantifiers were detected ($z_s = (-0.38, -0.4), p_s > 0.6$). Implicit learners (32 participants, $\beta = -2.43, z = -3.46, p = 0.0005$) and explicit incorrect learners (53 participants, $\beta = -2.33, z = -3.39, p = 0.0007$) were less accurate than explicit correct learners (5 participants). The small number of successful explicit learners shows that these quantifiers were particularly hard to learn. The effect of universal was also significant ($\beta = -0.60, z = -2.53, p = 0.01$). This suggests that in the case of conservativity, the learnability effect was robust and independent of the learning strategy that participants used. Fig. 8 illustrates the between-group differences in learning.

Results of the exploratory analyses show that accuracy is modulated by the quantifier condition, supporting the learnability hypothesis. Crucially, the "universal advantage" persists even when filtering for implicit and explicit incorrect learners, as we did not find any significant interactions.

4. Discussion

This study aimed to test the learnability hypothesis for three semantic universals in the domain of quantification. Under the learnability hypothesis, participants would learn quantifiers satisfying the universals of convexity (monotonicity), quantity, and conservativity more easily than quantifiers that do not satisfy these universals. These predictions bore out. We found a robust effect of learnability in all universals.

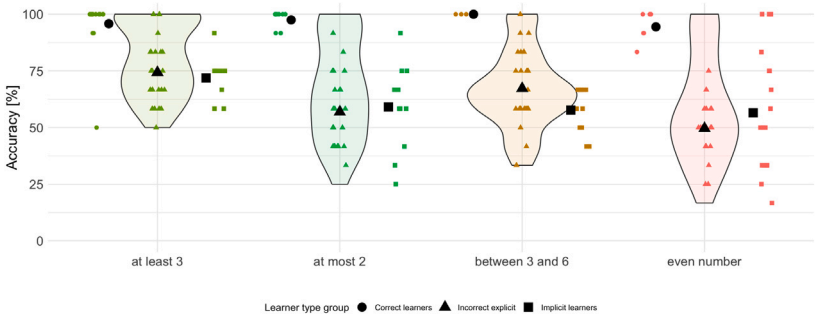

Fig. 6. Results of the exploratory analysis for monotonicity universal. The black shapes indicate the learner types mean, and the colored shapes the individual participants means. The explicit correct learners group (dot) has the highest accuracy across different quantifiers, while the accuracy of the two remaining groups (triangle, square) does not systematically vary.

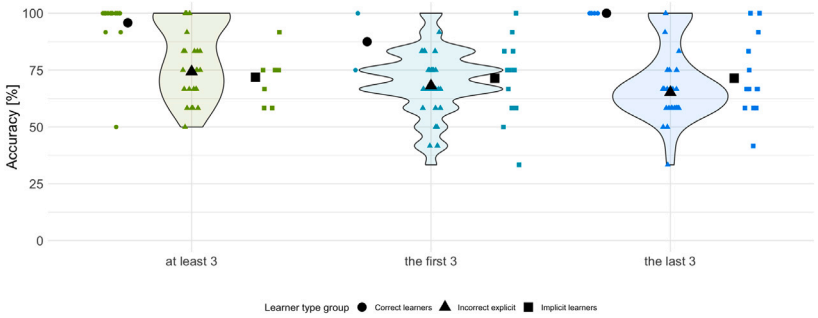

Fig. 7. Results of the exploratory analysis for quantity universal. The black shapes indicate the learner types mean, and the colored shapes the individual participants means. The explicit correct learners group (dot) has the highest accuracy across different quantifiers, while the accuracy of the two remaining groups (triangle, square) does not systematically vary.

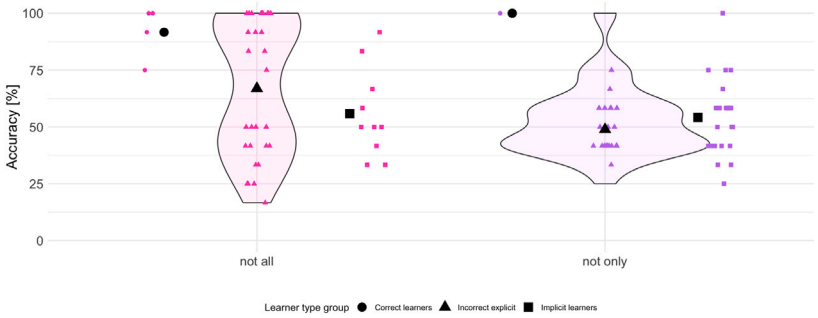

Fig. 8. Results of the exploratory analysis for conservativity universal. The black shapes indicate the learner types mean, and the colored shapes the individual participants means. The explicit correct learners group (dot) has the highest accuracy across different quantifiers, while the accuracy of the two remaining groups (triangle, square) does not systematically vary.

Participants achieved higher accuracy on the monotone quantifier *at least 3* than for *between 3 and 6* and *even number*, which are non-monotone. Moreover, they learned the quantitative quantifier better than the two non-quantitative quantifiers, and the conservative quantifier better than the non-conservative. We discuss these results in more detail below.

4.1. Monotonicity and convexity

Our results provide ample evidence of the role of learnability in shaping monotonicity/convexity universal. Previously, Chemla, Buccola, et al. (2019) found that adult participants learned monotone rules faster than non-convex rules. The difference between learning monotone rules and non-monotone rules was not significant. We added to this result by showing the monotonicity contrast. In our study, we found a difference between a monotone quantifier (*at least 3*) and non-monotone quantifiers (*between 3 and 6*, *even number*). When controlling for the different strategies that participants might have used to learn the meaning of the new word, our analysis revealed a difference between non-monotone and non-convex quantifiers.

In comparison with Chemla, Buccola, et al. (2019), we found our experiment challenging. For example, they found that participants learned the most difficult, non-convex rules in an average of 91 trials. In contrast, the training part of our experiment consisted of 84 trials, and only around one-fifth of the participants learned the non-convex quantifier. Even with the monotone quantifiers, 49% of participants did not perform systematically above chance. We suspect that the experimental design used by Chemla, Buccola, et al. (2019), with a fixed number of objects, helped participants constrain the space of possible quantifiers. Hence, they learned faster than in our experiment. Altogether, our study shows that changes in the design (e.g., variation in the number of stimuli) can increase difficulty and reveal differences between monotone and non-monotone quantifiers.

Our results are also compatible with those of Steinert-Threlkeld and Szymanik (2019). They showed that neural networks learned monotone quantifiers faster than non-monotone quantifiers. Yet, because the non-convex quantifier *at least 6 or at most 2* in their study was a negation of the non-monotone but convex quantifier *between 3 and 5*, the neural network did not encode the difference between convex and non-convex quantifiers, and thus Steinert-Threlkeld and Szymanik (2019) did not test the convexity contrast. Our experiment allowed us to measure learning difficulties caused by a lack of monotonicity and convexity. We found that non-monotone quantifiers are already more difficult to learn than monotone ones; however, the difference between learning *even number* and *between 3 and 6* was also significant in analysis controlling for learning strategies. In Ramotowska (2022) experiment, the non-convexity quantifier put a larger burden on learning than in this experiment. They tested, however, the non-convex quantifier *at most 2 or at least 7*, which has not been attested in the corpus. We selected a non-convex quantifier that is more comparable to the convex quantifiers in terms of frequency (see Appendix A.2). This suggests that the previous findings might have been driven, to some degree, by the frequency of quantifiers rather than by their universal properties.

We also found that the monotone-increasing quantifier was easier to learn than the monotone-decreasing one (cf similar difference found by Ramotowska, 2022). While the learnability hypothesis does not predict this result, this finding is consistent with the literature on quantifier acquisition (Geurts et al., 2010; Katsos et al., 2016). Moreover, this finding can be attributed to participants' performance on verification rather than to the learnability of downward-entailing quantifiers. Firstly, the monotone decreasing quantifiers are typically associated with processing difficulties, as demonstrated in many quantified sentence verification task studies (Deschamps et al., 2015; Just & Carpenter, 1971; Schlotterbeck et al., 2020; Szymanik & Zajenkowski, 2013). Thus, one explanation of our result is that participants learned the monotone-decreasing quantifier but made mistakes during the verification task, leading to lower accuracy. Secondly, the monotone decreasing quantifiers have an empty set among their witness sets (Bott et al. (2019); cf. Aloni (2022) for an explanation referring to the neglect zero tendency). Bott et al. (2019) showed that verifying the quantifiers with empty witness sets is particularly difficult and results in more errors and longer reaction times than verifying non-empty set situations (cf. Bott et al. (2025) for similar results). In our experiment, the quantifiers *at most 2* and *even number* were empty-set quantifiers, meaning that the empty set was among their witness sets. A closer look at the accuracy in the testing block of the experiment supports the above tentative explanation of our findings, given the added complexity introduced by the empty witness-set. We found that participants' accuracy was lowest (50%) when the number of target objects was zero, for *at most 2* and *even numbers* (compared to an average of 87% for the other two quantifiers). This suggests that the empty set as a witness set caused difficulties.

In this experiment, we tested the effect of monotonicity and convexity on the learning of determiners. However, the origin of the cognitive bias toward convexity seems to have a more primary root than in human language preference. Chemla, Dautriche, et al. (2019) showed that baboons (*Papio papio*) learn convex rules more easily than non-convex rules. Convexity also plays an important role in learning concepts and categories. The conceptual space is organized so that objects that share properties are more likely to be labeled with a single category. Metaphorically speaking, the concepts should not have gaps. For example, a set containing husky, labrador, and chihuahua forms a convex category of dogs, and a set containing dog, cat, and hamster forms a convex category of pets. Adult participants, when presented with a new word in the artificial language and example category members (like husky, labrador, and chihuahua), will generalize this new word to other members of the category (like beagle) at the same category level (Xu & Tenenbaum, 2007). However, when participants are presented with a set of objects that do not form a convex category (e.g., a set containing dog and tree), they will not generalize the new category label to all objects falling under the broader category (e.g., living things), but rather treat the new word as homophony (Dautriche & Chemla, 2016). Similarly, linguistic behavior and preference for convex categories were observed in children (Dautriche et al., 2016). Moreover, Steinert-Threlkeld and Szymanik (2020) showed that the degree of convexity predicts the neural network's ability to learn color term categories (cf. Gärdenfors, 2000 a convexity universal for colors). In addition, evidence from neuroscience suggests that conceptual representations in the brain are driven by geometric constraints, such as betweenness and equidistance, which are sufficient for forming convex categories (Bellmund et al., 2018). Thus, convexity bias seems to be present in human and non-human subjects, possibly biologically determined, and as a result, affects a wide range of cognitive phenomena related to categorization and language.

To summarize, our findings support the learnability explanation of the monotonicity universal. However, they also suggest that other pressures might shape the ease of learning of particular quantifiers, such as whether they have the empty set among their witness sets.

4.2. Quantity

Our results support the learnability explanation of the quantity universal. We showed a similar difference in the learning of quantitative and non-quantitative quantifiers, as Steinert-Threlkeld and Szymanik (2019) demonstrated with a neural network model. However, in the case of quantity, the universal effect was less robust and did not appear significant when controlling for the learning strategies of participants. Moreover, in the present experiment, we did not observe the difference in learning between non-quantitative quantifiers. In contrast, Steinert-Threlkeld and Szymanik (2019) showed that the neural networks learned *the last 3* slower than *the first 3*. Ramotowska (2022) obtained a similar result for human participants. These findings could not be explained by universal properties. However, the disparity between our approach and that of Ramotowska (2022) could also be explained by different approaches to analyzing learning data. In particular, they observed differences between the non-quantitative quantifiers only at the beginning of training, whereas we analyzed the data only from the testing part of the experiment.

4.3. Conservativity

We also found support for the role of learnability in shaping the conservativity universal. The finding is consistent with the results of Hunter and Lidz (2013) and Knowlton et al. (2022) (contrary to Spenader & de Villiers, 2019, and Ramotowska, 2022). In particular, we found that *not only* was the most difficult to learn. Nonetheless, we found that both quantifiers are learnable to some extent (cf. Knowlton et al., 2022). While *not only* was much more difficult to learn than *not all*, still two participants learned this quantifier. In a similar experimental setup but without the partitive construction *of the*, which focuses on the restrictor of the quantifier, Ramotowska (2022) did not find a difference in learning of conservative and non-conservative quantifiers. Comparison of this to our results suggests that the partitive construction made learning of *not only* particularly difficult. One could argue that because in English, the partitive construction is ungrammatical with the quantifiers such as *(not) only* (see Zamparelli (2008) for semantic analysis of partitive), participants (almost) never considered the non-conservative quantifiers as a potential meaning of *gleeb*. In other words, the difference in learning of *not only* and *not all* would be due to imposed grammatical constraints and not learning difficulties of non-conservative quantifiers *per se*. This would be in line with the structural account of conservativity (Romoli, 2015), according to which the conservativity universal is not a constraint on the lexicon and semantics of quantifiers, but a specific syntactic constraint on determiners (see Romoli (2015) for discussion of the Chierchia-Fox

hypothesis and Fox (2002)). In this view, the conservativity universal does not apply to adverbial quantifiers such as *(not) only* or comparative with *-er* morpheme, but only to the determiners. Our experiment cannot unequivocally confirm this hypothesis, but the comparison with the results of Ramotowska (2022) and Knowlton et al. (2022), who found that non-conservative verbs are not more difficult to learn than conservative ones, suggests that this hypothesis might be on the right track.

Our results contrast also with those of neural networks (Steinert-Threlkeld & Szymanik, 2019). Steinert-Threlkeld and Szymanik (2019) did not find any difference between the speed of learning of *not only* and *not all*. However, their result could be explained by the way in which quantifiers were encoded in the neural networks. This encoding did not emphasize the role of the restrictor, making the two quantifiers effectively symmetric. Humans, on the other hand, assign a special role to the restrictor. For example, Bott et al. (2025) showed that the empty restrictor cases (e.g., "All triangles are white" in case when there are no triangles) are detected fast and treated as presupposition violation, in contrast to the empty scope cases (e.g., "Fewer than three triangles are white" in case when all triangles are black), which are processed slowly. We believe that in our experiment, the use of *of the* construction focused the participants' attention on the restrictor and, therefore, made learning *not only* particularly difficult. This again suggests that conservativity universal might not constrain the semantics but might be rooted in the syntax-semantics interface.

4.4. Potential confounds and alternative explanations

Needless to say, adult participants are already affected by their experiences with their native languages, which creates potential confounds in our study. Although eliminating all these confounds is impossible, it is important to address them and consider whether factors other than the hypothesized semantic properties could have driven the results.

The objection could arise that participants in our experiment did not recognize the new word as a determiner and solved the task by learning a rule. To control for this confound, we used the partitive construction "of the", which should force the determiner interpretation of the new word. In addition, we asked participants to indicate what they thought *gleeb* meant. We asked two native English-speaking consultants to classify the answers as determiners or non-determiners. We found that 69% of the participants in our study recognized *gleeb* as a determiner, 24% of participants did not know what *gleeb* meant, and only 7% of participants had a non-determiner interpretation of the new word. This suggests that our experimental manipulation was successful and that participants learned *gleeb* as a determiner.

Another potential confounding factor is the frequency and familiarity of concepts. Even if participants considered the meaning of *gleeb* to be a determiner, they may not have learned the intended quantifiers; instead, they may have guessed a lexicalized quantifier with a meaning close to the intended one and mapped that meaning onto the new word. The post-experiment questionnaire showed that participants often mentioned quantifiers such as *most*, *some*, *majority*, or *half*. The experimental design addressed this confound by selecting stimuli so that participants could not achieve 100% accuracy in learning different intended quantifiers (see Appendix A.1 for more details). Thus, even when using the mapping strategy, participants would not achieve the same level of accuracy as if they had learned the correct meaning of *gleeb*. The validity of the quantifier choice was also confirmed empirically. In the additional analyses, we showed that across different universals, participants who recognized *gleeb* as a determiner and correctly learned its meaning achieved higher accuracy than those who learned other than intended determiners.

Our experimental paradigm imposed specific constraints on the type of learning strategy that participants could have applied. We conclude that, as intended, most of the participants used a mapping strategy. We base this claim on the fact that most participants recognized *gleeb* as a determiner and, in their typed responses, attempted to spell out its meaning as corresponding to one of the determiners in English, though many did so incorrectly. The most dominant incorrectly mentioned determiners were *most, the majority, some, all, none, more (than)*. Many participants mentioned more than one determiner, suggesting that they might have considered different meaning hypotheses. Some participants explicitly indicated that they tried to map between *gleeb* and known words (e.g., "I have no idea! I tried out 'all', 'half' and numbers 1-4, then I tried fractions and I still couldn't work it out! Sorry" or "I do not know, any ideas I had seemed to be proven wrong, such as half, all, none, some, etc."). However, based on the collected data, we were not in a position to definitively conclude which strategies participants have used.

Still, it is important to note that we found a significant effect of the universal property, even after controlling for learner groups. Furthermore, the lack of interaction between the universal and learner group demonstrates that the universal-satisfying learning advantage is not an artifact of an explicit mapping strategy. On the contrary, while the mapping strategy might bring participants' accuracy to the ceiling, the inherent learnability advantage of universal quantifiers is also observed in the groups that failed to explicitly map the concept.

Finally, we need to consider the limitations of our experimental paradigm in relation to real-world language learning. We acknowledge that our learning task does not resemble the type of learning that humans experience during language acquisition. In particular, children can learn mappings between words and semantic representations from only positive data (Hsu et al., 2013). In a well-controlled experiment with two response options, presenting an equal number of positive and negative instances of the quantifier aims to eliminate potential response biases (e.g., participants learning to use the more frequent option). Thus, the input data in this experiment has not reflected the distribution of data in real life. Nonetheless, the learning in our experiment resembles reinforcement learning, a well-known model of learning in cognitive psychology (Sutton et al., 2018), and is observed in a number of other cognitive tasks (e.g., Frank et al., 2004; Miletić et al., 2021; van de Vijver & Ligneul, 2020). Moreover, learning a new word as mapping could be conceptualized in the "Language of Thought" (LoT) framework as a process of finding the correct "mental formula" that corresponds to this word. Universal-satisfying quantifiers have simpler formulas in the human mental lexicon than non-universal ones (see van de Pol et al., 2023). As a result, they are more likely to be selected and tested as candidate meanings. In this framework, the learning bias toward monotonicity, quantity, and conservativity means that the hypothesis "a new word is a determiner satisfying universals" will be tested and confirmed much faster than "a new word is a determiner not satisfying universals". The data show that at least some learning can be described this way.

Further studies could explore the learnability hypothesis by fitting different learning models or manipulating different biases in the learning data (e.g., learning only from positive evidence cf. Hsu et al., 2013). Moreover, further studies are needed to better understand the learning process in our paradigm. For example, the interaction between the mapping strategy and universals could be further explored by giving participants a complex determiner to learn (e.g., *gleeb pra feezda*).[12] In our experiment, participants had to map known quantifiers into a single word meaning. Because of that, they might have constrained their search to lexicalized determiners (which are also universal-satisfying determiners). The use of a mapping strategy for complex phrases would not carry such a bias.

[12] We thank Claire Rong for this suggestion.

4.5. Learnability vs. other pressures

While this paper focuses on an empirical test of the learnability hypothesis, we must consider that learnability may stem from a broader general simplicity bias. A long tradition in cognitive psychology suggests that concepts definable with fewer logical operators are easier to acquire (Feldman, 2000; Shepard et al., 1961). Viewed through this lens, quantifier universals may simply reflect logical economy: quantifiers that violate convexity, quantity, or conservativity are more complex and, consequently, harder to learn. Formal analyses support this, showing that universal-satisfying quantifiers have shorter descriptions in a logical "language of thought" (van de Pol et al., 2023). However, mathematical simplicity is not a sufficient explanation on its own. For instance, under LoT-complexity *not all* and *not only* are equally simple, but they differ significantly in practical learnability. Such discrepancies highlight that theoretical complexity metrics may differ from human performance; empirical data are required to determine if the human cognitive system truly favors these universal properties in realistic learning scenarios.

Along with the simplicity bias, languages are shaped by multiple other pressures originating from cognitive, communicative, and historical constraints. Our study zoomed in on a cognitive learning bias, holding other factors constant. In the real world, however, those other factors also operate. One significant factor that shapes languages is informativeness. For example, the informativeness of the word predicts its length (Piantadosi et al., 2011), as short words tend to be more informative. Carr et al. (2020) proposed that, in addition to learnability and simplicity pressures, languages are shaped by informativeness and in this way, satisfy their communication goals. The communicative efficiency might favor certain meanings because they are more informative or useful in conversation, regardless of how easy they are to learn. Our study did not test communicative aspects, as each participant learned in isolation without using *gleeb* to communicate. However, according to the account of communicative efficiency, natural languages maintain the most optimal balance between informativeness and simplicity, which in turn facilitates communication (Kemp et al., 2018; Regier et al., 2015). For example, the complexity-informativeness trade-off shapes semantic categories, such as kinship terms (Kemp & Regier, 2012), animal and container names (Xu et al., 2016; Zaslavsky et al., 2020), color terms (Zaslavsky et al., 2018), person pronoun system (Zaslavsky et al., 2021), indefinite pronouns (Denić et al., 2020), and number systems (Denić & Szymanik, 2024; Xu et al., 2020). It is plausible that quantifier meanings that become lexicalized are also those that strike a balance between learnability and communicative effectiveness (Steinert-Threlkeld, 2021). In the case of quantifier universals, there is a strong reason to think these factors converge. A non-conservative determiner, for instance, in addition to being hard to learn, might lead to pragmatic oddity. It requires tracking elements outside the restrictor, which makes it less straightforward for a listener to interpret. Similarly, a non-quantitative determiner (like *the first 3*) might be less useful than a numerical one (like *at least 3*), because it is applicable in fewer situations. Future studies could investigate scenarios where learnability and communicative utility might conflict, or examine how learning biases manifest in actual child language acquisition of quantifiers (where pragmatics and input frequency come into play).

Beyond communication, other pressures have been proposed to contribute to linguistic universals. These include biases from the structure of the environment and cognition: for example, how our perceptual systems segment the world, how our memory constraints favor simpler categories, how cultural evolution filters language over generations, etc. In the domain of color naming and spatial terms, for example, environmental and perceptual factors clearly play a role in shaping universals (Culbertson et al., 2020; Regier et al., 2016). For quantifiers, one could imagine that constraints on the number representations (e.g., precise or approximate, see more in Dehaene, 1997) might influence which quantifiers get lexicalized. Recent works (e.g., Culbertson & Kirby, 2016) explore how iterated learning and cultural transmission can lead to simplification and alignment of linguistic systems. Community-level factors (such as population size or interaction networks) might also affect how complex a lexicon grows, with some evidence that smaller communities maintain more redundancy and simplicity in language (e.g., languages converging under contact tend to favor simpler, more regular structures Atkinson et al., 2018; Raviv et al., 2019). While our experiment does not directly address these macro-level forces, the support for the learnability hypothesis is consistent with other explanations. Instead, they complement a growing body of evidence from multiple methodologies indicating that quantifier universals have a cognitive basis. We emphasize a nuanced view: learnability is likely one explanatory factor among several, rather than the sole determinant of quantifier universals. In this view, the emergence of universals would be overdetermined, with learnability, functional communication needs, and historical-cultural dynamics all contributing in the same direction. In a counterfactual scenario, if our results had not revealed learning differences, we would have to turn toward other explanatory factors – perhaps the communicative or pragmatic efficiency. Instead, positive results indicate that we can attribute a significant portion of the explanation to learning biases. But even with this confirmation, we see learnability as one part of a larger explanatory framework.

In conclusion, our findings lend considerable weight to the idea that cognitive learnability has shaped the semantic landscape of quantifiers across languages. Prior to this work, the learnability hypothesis for quantifier universals was attractive but rested on relatively narrow evidence – some small-scale experiments and modeling efforts that each addressed parts of the issue. Our contribution is to add a critical piece of evidence from large-scale human experimentation supporting the learnability hypothesis. We directly show that adult human learners find quantifiers with "universal" properties easier to learn across all three major proposed universals under controlled conditions. This bridges a gap between theoretical speculation and empirical observation and provides direct empirical validation for the learnability hypothesis, which posits that these semantic universals are not arbitrary but arise from the way humans learn concepts. Learnability, in effect, acts as a filter through which meanings can become entrenched as words. However, we view this not as the sole filter but as one that likely works in tandem with others. By demonstrating how and why ease of learning correlates with linguistic universals, we move closer to a comprehensive understanding of why languages universally prefer certain meanings: those meanings align well with our minds. Our study adds crucial evidence to a multidisciplinary puzzle, supporting the notion that the universal constraints on quantifier meanings reflect, at least in part, fundamental constraints on human learning and cognition.

CRediT authorship contribution statement

Sonia Ramotowska: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Leendert van Maanen:** Writing – review & editing, Validation, Supervision, Methodology, Conceptualization. **Jakub Szymanik:** Writing – review & editing, Validation, Supervision, Resources, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Sonia Ramotowska was supported by the "Nothing is Logical" project (the Dutch Research Council (NWO) OC grant 406.21.CTW.023), as well as by the French National Research Agency (ANR) grants "Communicative efficiency, cognitive constraints and lexical meaning" (ANR-23-CE28-0016) and "FrontCog" (ANR-17-EURE-0017).

We would like to warmly thank the attendees of the "Experimental Approaches to Language Universals in Structure and Meaning" workshop at ESSLLI 2021, the "Internal and External Pressures Shaping Language" workshop at ESSLLI 2023, the SALC9 workshop "Explaining Semantic Universals," and the LINGUAE seminar (in particular Claire Rong and Zhuoye Zhao) at the Institute Jean Nicod for their helpful comments on this project. We are grateful to Dean McHugh, Patrick Elliott, and Todd Snider for classifying the typed responses from this and the pilot experiments. We also thank Erwin Schuijtvlot for his help in generating the stimuli for the experiment.

Appendix

A.1. Test stimuli

The test stimuli consisted of 12 items. For each item, we decided whether it is true or false given each of the most frequent quantifiers: *all*, *most*, *some*, *none*, and *at least half* as a hypothetical meaning of *gleeb*. Next, we computed the probability of giving a true response for each item based on the truth value assignments for the competing quantifiers. To give an example, for *at least 3* a sequence consisting of 7 objects: target, target filler, filler, target. target, target would be judged as true for *most*, *some*, and *at least half*, and false for *all* and *none*, so its probability of being true was 3 out of 5, meaning 0.6. We sample the input data covering all possible ranges (between 0.2 and 0.8, note that *some* and *none* are mutually exclusive, so given this set of quantifiers, the probability of *gleeb* being true could never be 0 or 1).

A.2. Quantifier frequency

Our experiment tested adult native English speakers who had already acquired the language. The frequency of quantifier usage could affect the ease of learning because it is easier to come up with the more frequent quantifiers for the meaning of the new word. To control for this confound, we checked the frequency of quantifiers used in our experiment. Frequencies were estimated using the LINDAT/CLARIAH-CZ implementation of the British National Corpus dependency treebank (total token count: 5,492,775 as of December 2025). To ensure we captured only genuine quantifier usage, we used PMLTQ queries with strict syntactic adjacency constraints (e.g., requiring the quantifier to precede the noun phrase immediately). For *not only*, we explicitly excluded constructions containing *but* to distinguish the quantifier usage from coordinate conjunctions.

We measured the frequency of the exact quantifiers that participants had to infer, as well as similar quantifier constructions. Table 2 summarizes the frequencies of occurrence, the exact quantifier from our study, and the quantifier construction like, DET (NUM), where DET is, for example, *at least*, NUM is the number in numerical quantifiers (e.g., 1, 2, 3, . . .). We used this as a more general measure.

We included two standard measures of frequency: frequency per million words (fpmw) and *Zipf* value (van Heuven et al., 2014). We used *Zipf* values (van Heuven et al., 2014; Zipf, 1949), a standardized scale that gives an intuitive interpretation of frequencies, to compare the frequencies of quantifiers used in this experiment. The *Zipf* values are calculated as a logarithm with base 10 of frequency per million words times 1000 (log10(fpmw*1000)). The *Zipf* values of 1 to 3 correspond to low-frequency words, and *Zipf* values between 4 and 7 to high-frequency words. Most of the quantifiers in our study had *Zipf* values between 4 and 5 *Zipf* values, meaning that they are rather frequent. Two quantifiers, *between 3 and 6* and *even number*, had *Zipf* values around 2, which classifies them as low-frequency words.

Table 2
Frequency table. For all numerical quantifiers, we searched numerical information as a word (e.g., three) or Arabic number (e.g., 3). The columns fpmw show the frequency per million words (and row frequency from the COCA corpus in the brackets) and the columns *Zipf* value show the word frequency computed as log10(fpmw*1,000).

Universal/Type	Quantifier	Raw Count	FPMW	Zipf Value
Conservativity	Not only	2994	545.08	5.74
	Not all	1316	239.59	5.38
Monotonicity	At least m[a]	3330	606.25	5.78
	At most m[a]	938	170.77	5.23
	At least 3	102	18.57	4.27
	At most 2	13	2.37	3.37
	Between m and k	1617	294.39	5.47
	Between 3 and 6	1	0.18	2.26
Convexity	Even number of	1	0.18	2.26
Quantity	The first n	5544	1009.33	6.00
	The last n	2035	370.49	5.57
	The first 3	392	71.37	4.85
	The last 3	189	34.41	4.54

[a] We also included counts of *more than m* and *fewer than n*.

Overall, all quantifiers used in our study were attested in the corpus. To test their effect on the participants' accuracy, we ran a logistic mixed-effects model with accuracy as the dependent variable and normalized Zipf value (of the exact quantifiers used in the study) and universal (satisfying vs. not-satisfying, contrast coded) as predictors (and by-subject random intercept). We conducted two analyses, in the first *an even number, the first/last 3* and *not only* were encoded as universal not-satisfying, and in the second one also *between 3 and 6* was also encoded as universal not-satisfying.

The first analysis showed only a significant effect of universal ($\beta = 0.41; p < 0.001$). The second analysis showed a significant effect of universal ($\beta = 0.48; p < 0.001$) and a significant effect of frequency ($\beta = 0.11; p = 0.04$). Overall, this suggests that the frequencies of quantifiers affected learnability; however, this effect was smaller than the effect of the universal.

A.3. Analysis without exclusions based on secondary task

Below we report the regression model analysis conducted on the whole sample of participants without excluding participants who did not meet the criteria of the secondary task. The results of these analyses fully aligned with the results after exclusion.

A.3.1. Monotonicity and convexity

The mean accuracy of the whole sample of participants was: *at least 3* (81%), *at most 2* (65.7%), *between 3 and 6* (65.8%), and *even number* (59.9%).

The mixed effect logistic regression analysis of accuracy with quantifiers *at least 3*, *at most 2*, *between 3 and 6*, and *even number* as predictors, followed by *post hoc* pairwise comparisons, led to the same result as the analysis of the subsample. The accuracy for *at least 3* was significantly higher than the accuracy for *at most* ($\beta = 0.90; p < 0.0007$), *between 3 and 6* ($\beta = 0.94; p = 0.0004$) and *even number* ($\beta = 1.19; p < 0.0001$). In contrast, there was no significant difference between *at most 2* and *between 3 and 6* ($\beta = 0.03; p > 0.99$), and between *at most 2* and *even number* ($\beta = 0.29; p = 0.57$), or between *even number* and *between 3 and 6* ($\beta = 0.25; p = 0.67$).

A.3.2. Quantity

The mean accuracy of the whole sample of participants was *at least 3* (81%), *the first 3* (69.8%), and *the last 3* (69.5%). Similarly, as for monotonicity, also for quantitativity, the analysis of the accuracy of the whole sample of participants led to the same result. We found a

significant difference between *at least 3* and *the first 3* ($\beta = 0.66; p = 0.0007$), and between *at least 3* and *the last 3* ($\beta = 0.67; p = 0.0006$), but not between *the first 3* and *the last 3* ($\beta = 0.01; p > 0.99$).

A.3.3. Conservativity

For conservativity as well, the pattern of results did not change when we included all participants in the analysis ($\beta = -0.71; p = 0.0018$), meaning that the accuracy was lower for *not only* (52.8%) than for *not all* (66%).

Data availability

The data associated with this publication are available at OSF: https://osf.io/gv25w.

References

Aarts, A. A., Anderson, J. E., Anderson, C. J., Attridge, P. R., Attwood, A., Axt, J., Babel, M., Bahník, Š., Baranski, E., Barnett-Cowan, M., Bartmess, E., Beer, J., Bell, R., Bentley, H., Beyan, L., Binion, G., Borsboom, D., Bosch, A., Bosco, F. A., ... Zuni, K. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), http://dx.doi.org/10.1126/science.aac4716.

Aloni, M. (2022). Logic and conversation: the case of free choice. *Semantics and Pragmatics*, *15*, 5–EA. http://dx.doi.org/10.3765/sp.15.5.

Atkinson, M., Smith, K., & Kirby, S. (2018). Adult learning and language simplification. *Cognitive Science*, *42*(8), 2818–2854. http://dx.doi.org/10.1111/cogs.12686.

Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy*, *4*(2), 159–219. http://dx.doi.org/10.1007/BF00350139.

Bellmund, J. L., Gärdenfors, P., Moser, E. I., & Doeller, C. F. (2018). Navigating cognition: Spatial codes for human thinking. *Science*, *362*(6415), eaat6766. http://dx.doi.org/10.1126/science.aat6766.

Bott, O., Schlotterbeck, F., & Klein, U. (2019). Empty-set effects in quantifier interpretation. *Journal of Semantics*, *36*(1), 99–163. http://dx.doi.org/10.1093/jos/ffy015.

Bott, O., Schlotterbeck, F., Klochowicz, T., Ramotowska, S., & Aloni, M. (2025). Neglect-zero effects in the interpretation of quantifiers and disjunction. *Proceedings of Sinn Und Bedeutung*, *29*, 214–232. http://dx.doi.org/10.18148/sub/2025.v29.1206.

Carcassi, F., Schouwstra, M., & Kirby, S. (2019). The evolution of adjectival monotonicity. In M. Espinal, E. Castroviejo, M. Leonetti, L. McNally, & C. Real-Puigdollers (Eds.), *Proceedings of sinn und bedeutung 23*: *Vol. 1*, (pp. 219–230). Bellaterra (Cerdanyola del Vallès): Universitat Autònoma de Barcelona, http://dx.doi.org/10.18148/sub/2019.v23i1.512.

Carcassi, F., Steinert-Threlkeld, S., & Szymanik, J. (2021). The emergence of monotone quantifiers via iterated learning. *Cognitive Science*, *45*, Article e13027. http://dx.doi.org/10.1111/cogs.13027.

Carr, J. W., Smith, K., Culbertson, J., & Kirby, S. (2020). Simplicity and informativeness in semantic category systems. *Cognition*, *202*, Article 104289. http://dx.doi.org/10.1016/j.cognition.2020.104289.

Chemla, E., Buccola, B., & Dautriche, I. (2019). Connecting content and logical words. *Journal of Semantics*, *36*(3), 531–547. http://dx.doi.org/10.1093/jos/ffz001.

Chemla, E., Dautriche, I., Buccola, B., & Fagot, J. (2019). Constraints on the lexicons of human languages have cognitive roots present in baboons (Papio papio). *Proceedings of the National Academy of Sciences of the United States of America*, *116*(30), 14926–14930. http://dx.doi.org/10.1073/pnas.1907023116.

Culbertson, J., & Kirby, S. (2016). Simplicity and specificity in language: Domain-general biases have domain-specific effects. *Frontiers in Psychology*, *6*, 1964. http://dx.doi.org/10.3389/fpsyg.2015.01964.

Culbertson, J., Schouwstra, M., & Kirby, S. (2020). From the world to word order: Deriving biases in noun phrase order from statistical properties of the world. *Language*, *96*(3), 696–717. http://dx.doi.org/10.1353/lan.2020.0045.

Dautriche, I., & Chemla, E. (2016). What homophones say about words. *PLoS One*, *11*(9), Article e0162176. http://dx.doi.org/10.1371/journal.pone.0162176.

Dautriche, I., Chemla, E., & Christophe, A. (2016). Word learning: Homophony and the distribution of learning exemplars. *Language Learning and Development*, *12*(3), 231–251. http://dx.doi.org/10.1080/15475441.2015.1127163.

Dehaene, S. (1997). *The Number Sense: How the mind creates mathematics*. New York: Oxford University Press.

Denić, M., Steinert-Threlkeld, S., & Szymanik, J. (2020). Complexity/informativeness trade-off in the domain of indefinite pronouns. In *Semantics and linguistic theory* (pp. 166–184). http://dx.doi.org/10.3765/salt.v30i0.4811.

Denić, M., & Szymanik, J. (2024). Recursive numeral systems optimize the trade-off between lexicon size and average morphosyntactic complexity. *Cognitive Science*, *48*(3), Article e13424. http://dx.doi.org/10.1111/cogs.13424.

Deschamps, I., Agmon, G., Loewenstein, Y., & Grodzinsky, Y. (2015). The processing of polar quantifiers, and numerosity perception. *Cognition*, *143*, 115–128. http://dx.doi.org/10.1016/j.cognition.2015.06.006.

Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature*, *407*(6804), 630–633. http://dx.doi.org/10.1038/35036586.

Fox, D. (2002). Antecedent-contained deletion and the copy theory of movement. *Linguistic Inquiry*, *33*(1), 63–96. http://dx.doi.org/10.1162/002438902317382189.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998. http://dx.doi.org/10.1126/science.1218633.

Frank, M. J., Seeberger, L. C., & O'reilly, R. C. (2004). By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science*, *306*(5703), 1940–1943. http://dx.doi.org/10.1126/science.1102941.

Gamer, M., Lemon, J., & Singh, I. F. P. (2019). irr: Various coefficients of interrater reliability and agreement. R package version 0.84.1, Retrieved from https://CRAN.R-project.org/package=irr.

Gärdenfors, P. (2000). *Conceptual spaces: the geometry of thought*. Cambridge: MIT Press, http://dx.doi.org/10.7551/mitpress/2076.001.0001.

Geurts, B., Katsos, N., Cummins, C., Moons, J., & Noordman, L. (2010). Scalar quantifiers: Logic, acquisition, and processing. *Language and Cognitive Processes*, *25*(1), 130–148. http://dx.doi.org/10.1080/01690960902955010.

Greenberg, J. H. (1978). Generalizations about numeral systems. In J. H. Greenberg (Ed.), *Universals of human language* (pp. 250–295). Stanford: Stanford University Press.

Hockett, C. F. (1963). The problem of universals in language. In J. H. Greenberg (Ed.), *Universals of language* (pp. 1–29). Cambridge, MA: MIT Press.

Hsu, A. S., Chater, N., & Vitányi, P. (2013). Language learning from positive evidence, reconsidered: A simplicity-based approach. *Topics in Cognitive Science*, *5*(1), 35–55.

Hunter, T., & Lidz, J. (2013). Conservativity and learnability of determiners. *Journal of Semantics*, *30*(3), 315–334. http://dx.doi.org/10.1093/jos/ffs014.

Hyman, L. M. (2008). Universals in phonology. *Linguistic Review*, *25*(1–2), 83–137. http://dx.doi.org/10.1515/TLIR.2008.003.

Just, M. A., & Carpenter, P. A. (1971). Comprehension of negation with quantification. *Journal of Verbal Learning and Verbal Behavior*, *10*(3), 244–253. http://dx.doi.org/10.1016/S0022-5371(71)80051-8.

Katsos, N., Cummins, C., Ezeizabarrena, M.-J., Gavarró, A., Kuvač Kraljević, J., Hrzica, G., Grohmann, K. K., Skordi, A., Jensen de López, K., Sundahl, L., et al. (2016). Cross-linguistic patterns in the acquisition of quantifiers. *Proceedings of the National Academy of Sciences*, *113*(33), 9244–9249. http://dx.doi.org/10.1073/pnas.1601341113.

Keenan, E. L., & Stavi, J. (1986). A semantic characterization of natural language determiners. *Linguistics and Philosophy*, *9*(3), 253–326. http://dx.doi.org/10.1007/BF00630273.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*(6084), 1049–1054. http://dx.doi.org/10.1126/science.1218811.

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, *4*(1), 109–128. http://dx.doi.org/10.1146/annurev-linguistics-011817-045406.

Knowlton, T., Trueswell, J., & Papafragou, A. (2022). New evidence for the unlearnability of non-conservative quantifiers. In M. Degano, T. Roberts, G. Sbardolini, & M. Schouwstra (Eds.), *The proceedings of the 23rd amsterdam colloquium*.

Kochari, A. R. (2019). Conducting web-based experiments for numerical cognition research. *Journal of Cognition*, *2*(1), 39. http://dx.doi.org/10.5334/joc.85.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), http://dx.doi.org/10.18637/jss.v082.i13.

Maldonado, M., & Culbertson, J. (2022). Person of interest: Experimental investigations into the learnability of person systems. *Linguistic Inquiry*, *53*(2), 295–336. http://dx.doi.org/10.1162/ling_a_00406.

Maldonado, M., Culbertson, J., & Uegaki, W. (2022). Learnability and constraints on the semantics of clause-embedding predicates. In *Proceedings of the annual meeting of the cognitive science society*: *Vol. 44*.

Miletić, S., Boag, R. J., Trutti, A. C., Stevenson, N., Forstmann, B. U., & Heathcote, A. (2021). A new model of decision processing in instrumental learning tasks. *Elife*, *10*, Article e63055. http://dx.doi.org/10.7554/eLife.63055.

Mostowski, A. (1957). On a generalization of quantifiers. *Fundamenta Mathematicae*, *44*(1), 12–36. http://dx.doi.org/10.2307/2964414.

Peters, S., & Westerståhl, D. (2008). *Quantifiers in language and logic* (pp. 1–552). New York: Oxford University Press, http://dx.doi.org/10.1093/acprof:oso/9780199291267.001.0001.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529. http://dx.doi.org/10.1073/pnas.1012551108.

van de Pol, I., Lodder, P., van Maanen, L., Steinert-Threlkeld, S., & Szymanik, J. (2023). Quantifiers satisfying semantic universals have shorter minimal description length. *Cognition*, *232*, Article 105150. http://dx.doi.org/10.1016/j.cognition.2022.105150.

Ramotowska, S. (2022). *Quantifying quantifier representations: Experimental studies, computational modeling, and individual differences*. University of Amsterdam.

Ramotowska, S., Marty, P., Van Maanen, L., & Sudo, Y. (2024). Some but not all speakers sometimes but not always derive scalar implicatures. In *Proceedings of the annual meeting of the cognitive science society*: *Vol. 46*.

Raviv, L., Meyer, A., & Lev-Ari, S. (2019). Larger communities create more systematic languages. *Proceedings of the Royal Society B*, *286*(1907), Article 20191262. http://dx.doi.org/10.1098/rspb.2019.1262.

Regier, T., Carstensen, A., & Kemp, C. (2016). Languages support efficient communication about the environment: Words for snow revisited. *PLoS One*, *11*(4), Article e0151138. http://dx.doi.org/10.1371/journal.pone.0151138.

Regier, T., Kemp, C., & Kay, P. (2015). Word meanings across languages support efficient communication. In *The handbook of language emergence* (pp. 237–263). Wiley Online Library.

Romoli, J. (2015). A structural account of conservativity. *Semantics-Syntax Interface*, *2*, 28–57.

Schlotterbeck, F., Ramotowska, S., van Maanen, L., & Szymanik, J. (2020). Representational complexity and pragmatics cause the monotonicity effect. In S. Denison, M. Mack, Y. Xu, & B. Armstrong (Eds.), *Proceedings of the 42nd annual meeting of the cognitive science society* (pp. 3398–3404). Cognitive Science Society.

Schnoebelen, T., & Kuperman, V. (2010). Using amazon mechanical turk for linguistic research. *Psihologija*, *43*(4), 441–464. http://dx.doi.org/10.2298/PSI1004441S.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications.. *Psychological Monographs: General and Applied*, *75*(13), 1. http://dx.doi.org/10.1037/h0093825.

Spenader, J., & de Villiers, J. (2019). Are conservative quantifiers easier to learn? Evidence from novel quantifier experiments. In J. J. Schlöder, D. McHugh, & F. Roelofsen (Eds.), *Proceedings of the 22nd amsterdam colloquium* (pp. 504–512). University of Amsterdam.

Steinert-Threlkeld, S. (2020). Quantifiers in natural language optimize the simplicity/informativeness trade-off. In J. J. Schlöder, D. McHugh, & F. Roelofsen (Eds.), *Proceedings of the 22nd amsterdam colloquium* (pp. 513–522). University of Amsterdam.

Steinert-Threlkeld, S. (2021). Quantifiers in natural language: Efficient communication and degrees of semantic universals. *Entropy*, *23*(10), http://dx.doi.org/10.3390/e23101335.

Steinert-Threlkeld, S., & Szymanik, J. (2019). Learnability and semantic universals. *Semantics and Pragmatics*, *12*(4), 1. http://dx.doi.org/10.3765/sp.12.4.

Steinert-Threlkeld, S., & Szymanik, J. (2020). Ease of learning explains semantic universals. *Cognition*, *195*, Article 104076. http://dx.doi.org/10.1016/j.cognition.2019.104076.

Sutton, R. S., Barto, A. G., et al. (2018). *Reinforcement learning: An introduction*: *Vol. 1, No. 1*, (2nd ed.). MIT press Cambridge.

Szymanik, J. (2016). *Quantifiers and cognition. logical and computational perspectives*. Springer.

Szymanik, J., & Kieraś, W. (2022). The semantically annotated corpus of polish quantificational expressions. *Language Resources and Evaluation*, *56*(3), 1057–1074. http://dx.doi.org/10.1007/s10579-022-09578-4.

Szymanik, J., & Zajenkowski, M. (2013). Monotonicity has only a relative effect on the complexity of quantifier verification. In M. Aloni, M. Franke, & F. Roelofsen (Eds.), *Proceedings of the 19th amsterdam colloquium* (pp. 219–225). University of Amsterdam.

van Benthem, J. F. (1986). *Studies in linguistics and philosophy*: *Vol. 29*, *Essays in logical semantics*. Dordrecht: Springer.

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: a new and improved word frequency database for British english. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. http://dx.doi.org/10.1080/17470218.2013.85.

van de Vijver, I., & Ligneul, R. (2020). Relevance of working memory for reinforcement learning in older adults varies with timescale of learning. *Aging, Neuropsychology, and Cognition*, *27*(5), 654–676. http://dx.doi.org/10.1080/13825585.2019.1664389.

Xu, Y., Liu, E., & Regier, T. (2020). Numeral systems across languages support efficient communication: From approximate numerosity to recursion. *Open Mind*, *4*, 57–70. http://dx.doi.org/10.1162/opmi_a_00034.

Xu, Y., Regier, T., & Malt, B. C. (2016). Historical semantic chaining and efficient communication: The case of container names. *Cognitive Science*, *40*(8), 2081–2094. http://dx.doi.org/10.1111/cogs.12312.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference.. *Psychological Review*, *114*(2), 245–272. http://dx.doi.org/10.1037/0033-295X.114.2.245.

Zamparelli, R. (2008). Dei ex machina: a note on plural/mass indefinite determiners. *Studia Linguistica*, *62*(3), 301–327. http://dx.doi.org/10.1111/j.1467-9582.2008.00149.x.

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, *115*(31), 7937–7942. http://dx.doi.org/10.1073/pnas.1800521115.

Zaslavsky, N., Maldonado, M., & Culbertson, J. (2021). Let's talk (efficiently) about us: Person systems achieve near-optimal compression. *Vol. 43*, In *Proceedings of the annual meeting of the cognitive science society*.

Zaslavsky, N., Regier, T., Tishby, N., & Kemp, C. (2020). Semantic categories of artifacts and animals reflect efficient coding. In *Proceedings of the society for computation in linguistics 2020* (pp. 80–81).

Zipf, G. (1949). *Human behaviour and the principle of least effort*. Addison-Wesley.