Finding the force: How children discern possibility and necessity modals

Anouk Dieuleveut[1], Annemarie van Dooren[2], Ailís Cournane[3] and Valentine Hacquard[4]

Corresponding author: Anouk Dieuleveut, anouk.dieuleveut@gmail.com
ORCID: 0000-0002-8633-8672

Abstract

This paper investigates when and how children figure out the force of modals: that possibility modals (e.g. *can/might*) express possibility, and necessity modals (e.g. *must/have to*), necessity. Modals raise a classic subset problem: given that necessity entails possibility, what prevents learners from hypothesizing possibility meanings for necessity modals? Three solutions to such subset problems can be found in the literature: the first is for learners to rely on downward-entailing environments (Gualmini and Schwarz 2009); the second is a bias for strong (here, necessity) meanings; the third is for learners to rely on pragmatic cues, stemming from the conversational context (Dieuleveut et al. 2019, Rasin and Aravind 2020). This paper assesses the viability of each of these solutions, by examining the modals used in speech to and by 2-year-old children, through a combination of corpus studies and experiments testing the guessability of modal force based on their context of use. Our results suggest that given the way modals are used in speech to children, the first solution is not viable, and the second unnecessary. Instead, we argue that the conversational context in which modals occur is highly informative as to their force, and sufficient, in principle, to sidestep the subset problem. Our child results further suggest an early mastery of possibility, but not necessity modals, and show no evidence for a necessity bias.

Keywords

Modals, Negation, Language acquisition, Corpus study, Human Simulation Paradigm experiment

Acknowledgments

This project is supported in part by NSF grant #BCS-1551628. We thank Alexander Williams, Jeff Lidz, Tyler Knowlton, Yu'an Yang, Anissa Zaitsu, Dan Goodhue, Rémi Varloot, the Modality Group, the Language Acquisition Lab at the University of Maryland and Florian Schwarz and two anonymous reviewers for helpful comments and suggestions, and our research assistants Monica Pizzo, Joon Lee, David Whitcomb, Madisen Fong and Avni Gulrajani.

[1] Université de Paris Cité, CNRS, LLF, Paris, France
[2] University of Groningen, Groningen, The Netherlands
[3] New York University, New York
[4] University of Maryland, College Park

Introduction

Modals are used to talk about possibilities and necessities, that is, non-actual states of affairs. This paper investigates how children figure out the force of the modals in their language: that words like *can, may* or *might* in (1a) express possibility, whereas words like *must, should* or *have to* in (1b) express necessity.

(1) a. You *can/may/might...* go this way. possibility (◊)
b. You *must/should/have to/...* go this way. necessity (□)

The experimental literature on children's modal comprehension suggests that they struggle with modal force until at least age 4: they tend to both accept possibility modals in necessity situations, and necessity modals in possibility situations (e.g., Noveck 2001; Ozturk and Papafragou 2015). Typically, these errors are attributed to reasoning difficulties: children over-accept possibility modals in necessity situations because of difficulties reasoning about when a stronger modal would be more appropriate (i.e., they have trouble with scalar implicatures); they accept necessity modals in possibility situations because of difficulties reasoning about open possibilities (Acredolo and Horobin 1987). Usually, these studies take for granted that children already know the underlying force of modals. However, children's difficulties could reflect a lack of knowledge of their underlying force. In this paper, we address more directly the questions of when and how children figure out modal force, by investigating modal talk to and by young children, with a corpus study using the Manchester Corpus of UK English (Theakston et al., 2001; CHILDES database, MacWhinney 2000), and four experiments based on a modified version of the *Human Simulation Paradigm* (Gillette et al. 1999), testing how well adult participants can guess the force of modals uttered by either children or their mothers from the conversational context alone.

Imagine a child who hears a new modal, *sig*: 'You sig go this way.' How can she determine whether *sig* expresses necessity or possibility? The modal's syntactic position, before a verbal complement, might help narrow candidate meanings to expressing some kind of modal meaning (in the spirit of Landau and Gleitman's 1985 syntactic bootstrapping hypothesis), but it cannot help distinguish force, since possibility and necessity modals can appear in all the same syntactic environments. Cues from the physical context are also bound to be limited, since modals express non-actual concepts, with few physical correlates (Landau and Gleitman 1985). To learn the force of modals, children might thus need to rely heavily on cues from the conversational context. But how informative is the context about modal force?

One issue that might make this mapping of modal form to force particularly challenging is that necessity entails possibility: If you *must* go this way, considering the range of passable routes, then you *can* go this way, given that same set of options. Likewise, if you *must* eat with your right hand, given the rules of etiquette, then those same rules imply you *can* eat with your right hand. So, if you think that *sig* means 'possible' but in fact it means 'necessary', it is unclear how you can discover that in fact, *sig* has a stronger meaning: in situations where a necessity modal is used,

a possibility statement is also systematically true. What then prevents learners from postulating possibility meanings for necessity modals like *must* or *have to*? This kind of *subset* or *entailment problem* arises whenever two words' meanings enter into a set/subset relationship, and has been discussed for content words, like *dog*/*animal* (*e.g.* Xu and Tenenbaum 2007), as well as quantifiers like *some*/*every* and numerals (e.g. Piantadosi 2011; Piantadosi et al. 2013; Rasin and Aravind 2020). In this paper, we focus on this issue for modals.

Different types of solutions have been proposed in the literature for how learners resolve or sidestep subset problems.[5] The first one is for them to rely on downward entailing environments, which reverse patterns of entailment, as Gualmini and Schwarz 2009 suggest as a general solution to subset problems. The second is a bias towards strong (here, necessity) meanings, in the spirit of Berwick (1985). The third one is for the conversational context in which modals occur to be rich enough for learners to infer their force, without having to rely on either downward entailing environments or a necessity bias (Dieuleveut et al. 2019; Rasin and Aravind 2020 for *every*).

According to the first solution, all that children need to solve the subset problem is to observe necessity modals in downward-entailing (DE) environments, for instance under negation, as these environments reverse patterns of entailment (*not possible* entails *not necessary*). If children hear 'You *don't have to* go this way', in a situation where it is clear that there are other ways to go, they should be able to infer that *have to* doesn't express possibility: if it did, its negation would mean *impossible*, and wouldn't allow for other ways. We will argue that this is not a viable solution for modals. First, our corpus results show that necessity modals rarely occur with negation, let alone in other DE-environments, in the actual input to children. Second, problems arise from the fact that scope relation between modals and negation are idiosyncratic. Necessity modals do not uniformly scope under negation: *have to* does, but *must* or *should* do not (Iatridou and Zeijlstra 2013). Third, instances where necessity modals do occur with negation are least informative about their force: our experimental results show that participants had the most difficulty guessing the force of necessity modals in negative contexts.

[5] In this paper, we put aside two other possibilities. First, children could learn force on the basis of explicit negative feedback: they would be corrected by adults when they produce necessity modals to describe mere possibilities. We leave a systematic investigation of whether children receive such feedback, but, anecdotally, we haven't encountered any. In their study on *every*, Rasin & Aravind (2021) report only one case of explicit negative feedback, out of 72 *every* uttered by children. Second, children could exploit the principle of contrast (Clark & MacWhinney 1987) (e.g., knowing a possibility modal could lead children to hypothesize a necessity meaning for the next modal they encounter). However, as we discuss in section 1.1, learners cannot expect two modals to necessarily express different forces, since several modals express the same force. If children heard necessity claims explicitly contrasted with possibility claims (e.g. "you can, but you don't have to"), they could infer that the former expresses possibility, and the latter necessity, based on logical compatibility. While we did not systematically examine the transcripts for such cases, we did extract all cases where two modal claims were related by the logical connectives *or, but* and *and*. We found only 85 such cases (out of 18,853 utterances) Among these, only three were possibly useful to figure out the force of possibility modal (e.g., "it might be in your box or it might be lost"), but we found no informative cases for necessity modals. This may be an underestimation since we did not look for the occurrence of modals across different utterances. However, this very low frequency makes it unlikely that children can learn from such contrasts.

If learners cannot clearly rely on DE-environments, they may need a bias towards necessity meanings. According to this second solution, children would assume necessity meanings by default, and revise their hypothesis only for possibility modals, when hearing them used in situations of non-necessity. This kind of solution, proposed for other instances of the *subset problem,*[6] has been criticized by many authors, both on conceptual (Gualmini and Schwarz 2009) and empirical grounds (Xu and Tenenbaum 2007; Musolino 2006; Piantadosi et al. 2011, 2013; Rasin and Aravind 2020; for a summary, see Musolino et al. 2019). But such a bias could be necessary in the case of modals: there may for instance be fewer visual cues about their meanings than for concrete objects or even quantifier meanings, since they express abstract concepts about the non-actual.

In this paper, we argue that a necessity bias may not be necessary, even in the case of modals,-and that the subset problem is in principle solvable based solely on cues stemming from the conversational context in which modals occur (Dieuleveut et al. 2019). Rasin and Aravind (2020) reach a similar conclusion for *every*: while truth-conditional evidence alone may not allow the child to block an existential meaning (e.g., *some*) for *every*, pragmatic evidence can play an important role in sidestepping it. Rasin and Aravind consider potential sources of truth-conditional evidence against an existential meaning: downward-entailing environments, non-monotonic environments, and environments where the existential quantifier cannot be used (NPI-licensing, *almost*, *nearly*, and exceptive *but*). They show that such cases are very rare in the input (1.75% of *every* utterances), but that informative pragmatic evidence for rejecting an existential meaning for *every* is systematically available. Indeed, often enough (17%) when *every* is used in questions or assertions, its existential counterpart is already part of the common ground (as determined by the two authors). Hence, if *every* had an existential meaning, its use would result in a trivial contribution (e.g., asking "is someone here?" or asserting "someone is here" in a context where someone has arrived). They conclude that, if children assume that speakers do not make vacuous contributions, they could infer that *every* does not have an existential meaning, and thus sidestep the subset problem without special learning principles.

In this study, we probe the role of context in a different way, by having naïve speakers guess the force of modals based on the context of use, expanding on Dieuleveut et al. 2019[7]. Our experimental results show that this conversational context is informative about modal force: for the most part, participants were able to accurately recover the force of both possibility and necessity modals from mere snippets of conversation. Thus, in principle, learners should be able

[6] Many variants of this idea can be found in the literature. The *Subset Principle* (Baker 1979; Pinker 1979; Dell 1981; Berwick 1985; Manzini and Wexler 1987; a.o.) was originally proposed for the acquisition of syntactic phenomena. Later on, the *Semantic Subset Principle* (SSP) was introduced by Crain and Thornton (1998) to account for *semantic* set/subset problems at the sentential level (see also Crain et al. 1994; Crain 2012).

[7] We were unaware of Rasin and Aravind's methodology when we started this project with Dieuleveut *et al*. 2019, and thus approached the problem for modals differently. We leave an application of Rasin and Aravind's methodology to modals for future research, though we suspect that it might not always be easy to tell whether a possibility is already part of the common ground when a necessity statement is made, as all kinds of possibilities may be open at any given point in a conversation, even if they may not be salient.

to figure out modal force based on cues from the conversational context alone, and solve the subset problem without having to rely on downward-entailing contexts nor on a necessity bias. But, while a necessity bias is in principle not necessary on the basis of the input, children may still make use of one in practice. However, we find no evidence for a necessity bias in young children's modal productions.

Our current understanding of young children's modal force use is limited. Comprehension studies (e.g. Hirst and Weil 1982, Byrnes and Duff 1989, Noveck 2001, Ozturk and Papafragou 2015, a.o.) tend to focus on older children. Corpus studies (e.g. Kuczaj and Maratsos 1975; Papafragou 1998; Cournane 2015a,b, 2021; van Dooren et al. 2017) tend to focus on modal flavor, and while they note when particular lexemes first appear in children's productions, to date, no study systematically examines modal force in naturalistic productions (but see Jeretič 2018, and Dieuleveut et al. 2019). In this paper, we provide the first large scale study of the development of modal force, by examining the modal production of twelve children between the ages of 2 and 3. Our corpus and experimental results on children's modals indicate an asymmetry in force acquisition. Children seem to master possibility modals early: at age 2, children use them frequently and productively, both with and without negation. And, they use them in an adult-like way: crucially, they do not use them in necessity situations. However, they seem to struggle with necessity modals. They produce these much less frequently, and often, in a non adult-like way: they use them in situations where adults would prefer possibility modals. If this difficulty with necessity modals persists into the preschool years, it could explain children's tendency in prior comprehension studies to both accept possibility modals in necessity contexts (they may lack a relevant stronger alternative), and necessity modals in possibility contexts (they may not be sure that these modals express necessity).

Together, our results from mothers' and children's productions seem to lead to a puzzle: if the conversational context is informative about both forces, why should children particularly struggle with necessity modals? The early advantage for possibility modals could be due to a combination of factors. First, possibility modals are more frequent than necessity modals in children's input (about 3:1). Second, situations in which possibility modals occur with negation seem to be particularly informative (e.g. prohibitions, impossibilities), while negation may be particularly misleading with necessity modals. Whatever the reason for children's difficulty with necessity modals, their successes with possibility modals and relative failures with necessity modals provide no evidence for a necessity bias. Given that a necessity bias is neither necessary, in view of the information available in the input, nor is it evidenced in children's productions, we suggest that it is dispensable, even in the case of modals.

The rest of this paper is structured as follows. In section 1, we provide some general background on modal force and its acquisition. We first give a brief overview of the semantics and pragmatics of modals in English and beyond, particularly as they relate to force, and discuss the possible learnability implications that these cross-linguistic considerations engender. We then turn to how modals interact with negation, and what this might entail for force acquisition. We then review the main relevant findings from the modal acquisition literature. In section 2, we present

our input study. We first provide a descriptive, quantitative assessment of the modals children hear: which modals occur and how often, and when they appear with negation and in other DE-environments. We then present three input-based experiments, which assess the general informativity of natural conversational contexts about modal force, by asking adult participants to guess a modal blanked out from a dialogue extracted from the corpus, following a modified version of the *Human Simulation Paradigm* (HSP) (Gillette et al. 1999). In Experiment 1, the blanked modal statement is presented in context (7 preceding lines of dialogue), in Experiment 2, it is presented without the dialogue, with only the target sentence, in Experiment 3, all content is removed, as the sentence is presented with content words replaced by nonce 'Jabberwocky' words. Our results show that the conversational context modals are used in is highly informative about both forces. We then probe what aspects of the conversational context might be helpful, and identify one feature in particular for root modals, namely the desirability of the prejacent. A fourth experiment confirms that necessity, but not possibility modals, are typically used with undesirable prejacents. In section 3, we turn to children's productions. We first provide a quantitative assessment of the modals they produce, and then present a fifth experiment, which assesses the extent to which children use their modals in an adult-like way, by asking adult participants to guess the force of modals used by children. Our results suggest that children master possibility modals early, but struggle with necessity modals. In section 4, we discuss implications of our findings for how modal force acquisition might unfold in English, and beyond.

1 Background

1.1 Modal force in English and beyond

English modals come in two main forces: possibility and necessity. This is standardly captured by treating modals as existential or universal quantifiers over possible worlds, following the modal logic tradition. Further force distinctions can be found: necessity modals can be split into strong (*must*) vs. weak (*should*) necessity (von Fintel and Iatridou 2008);[8] nouns (*slight possibility*) and adjectives (*likely*) can encode even finer-grained strength distinctions. Here we focus on the main contrast between possibility and necessity modals and the learnability issues that it gives rise to.

Modals can be used to express different flavors of modality: epistemic modals (as in (2)) express possibilities and necessities given some evidence; deontic modals, possibilities and necessities given some relevant rules (as in (1)). We will use the term 'root' modality (Hoffmann 1966) for all non-epistemic flavors. This distinction will matter for us in that root modals tend to pattern together and differently from epistemic modals, in their interactions with scope-bearing elements, notably negation.

[8] The difference between weak and strong necessity is illustrated in the following example: 'Employees must wash their hands. Everyone else should.' (von Fintel and Iatridou 2008). Weak necessity modals are still treated as necessity modals, but quantify over a smaller domain than their strong counterparts.

(2) a. It might be raining. possibility (◊)
b. It must be raining. necessity (□)

In English, a modal always expresses the same force (possibility or necessity). However, it can be used for different flavors: '*Jo must draw'* can express an epistemic necessity ('Jo is likely to draw'), or a teleological, bouletic, or deontic necessity ('Jo needs/wants/is required to draw'). This is captured in the classical Kratzerian framework (Kratzer 1981, 1991) by having modals be lexically specified for force, but not for flavor. Flavor gets determined by conversational backgrounds which specify the set of worlds that the modal quantifies over, as the lexical entries, slightly modified from Kratzer (1991), illustrate in (3).

(3) For any world w, conversational background f:[9]
a. $[[\text{can}]]^{w,f}=\lambda q_{<st>}.\exists w' \in \cap f(w)\colon q(w')=1$
b. $[[\text{must}]]^{w,f}=\lambda q_{<st>}.\forall w' \in \cap f(w)\colon q(w')=1$

According to Horn (1972), modals form scales ($<\text{can}_{\text{deontic}}$, have $\text{to}_{\text{deontic}}>$,$<\text{might}_{\text{epi}}$, $\text{must}_{\text{epi}}>\ldots$),[10] and as such, they give rise to scalar implicatures (SI). The use of (1a), for instance, can implicate that you don't have to go this way, the use of (2a), that it doesn't have to be raining. In the Gricean tradition (1975), this implicature arises from the assumption that the speaker is trying to be maximally informative, but is not in a position to assert the relevant stronger statement in (2b). Speakers should prefer to use *must p* whenever they believe it to be true and relevant: listeners can then infer from the fact that the speaker did not chose the stronger (more informative) sentence that she must not believe it.

In Indo-European languages like English, possibility and necessity duals are common. However, various languages seem to lack such pairs. Instead, the same 'variable force' modals can be used in situations where English speakers would either use a possibility, or a necessity modal. Analyses vary in how to capture these variable force behaviors (see Yanovich 2013 for a summary). In St'˝at'imcets and Washo, modals have been analyzed as underlyingly necessity (universal) modals, which can be weakened by contextually restricting their domain of quantification (Rullmann et al. 2008, Bochnak 2015). In Nez Perce, the modal *o'qa* has been analyzed as a possibility (existential) modal, whose apparent variable force is due to the lack of a

[9] We ignore the ordering source here, which can derive further gradability and flavor differences amongst root modals.
[10] Logical entailment relations hold within flavor only: for example, epistemic necessity (e.g. 'given what we know, he must be upstairs.') does not entail deontic possibility (e.g. 'given the rules, he can be upstairs.'). Horn scales are thus defined within a flavor. Because of flavor variability, this means that the same lexeme can appear in different scales. We leave aside debates about ability modals, often argued to have no necessity counterpart (Horn 1972, Hackl 1998) (*e.g.* 'Jo **can** speak German, in fact, he has to' leads to oddity, or forces a switch in flavor interpretation). It is also argued that ability modals do have duals, *compulsion* modals, which are just extremely infrequent (*e.g.* 'I have to sneeze') (Mandelkern at al. 2015).

lexicalized stronger necessity dual in the language: *o'qa* does not belong to a Horn-scale, therefore its use is never associated with a scalar implicature (Deal 2011). Gitksan *ima* is similarly analyzed as a possibility modal (Matthewson 2013; Peterson 2010).[11]

Turning back to our learning problem, the range of cross-linguistic variation we find suggests that there may be few constraints on the space of hypotheses learners have to entertain for modals. They can't expect modals to come in duals, nor that their language must have a possibility modal, nor a necessity modal. And even in a language with duals like English, knowing the force of *one* modal doesn't guarantee that the next modal will express a different force, given that several lexemes can express the same force (e.g., *can, might* and *may*): children will thus have to figure out force for each modal anew.

One aspect of the English modal system that could indirectly help the learner is that speakers may refrain from using possibility modals in necessity situations, since necessity modals would be more informative. If the situations in which possibility modals are used never overlap with those in which necessity modals are used, this could help English learners distinguish possibility from necessity modals. However, the extent to which adults always choose to use necessity modals over possibility modals in necessity situations is not entirely clear. Speakers do not always aim for maximal informativity: other conversational principles intervene. Possibility modals can be used, for instance, to soften statements in a polite way: 'You *could* be a little more quiet' can be used as an order to be quiet, or 'It *might* be too late' to convey that it *is* too late (Searle 1975, Grice 1975, Austin 1975, Brown and Levinson, 1987, a.o.). Note that these politeness considerations are peculiar to modals, and do not arise, for instance, with quantifiers over individuals. If frequent enough, they could blur the distinction between possibility and necessity modals and be particularly misleading. One of our main goals here is to find out how clear the input is about the underlying force of modals in speech to children.

We now turn to the interaction of modals with negation, and discuss the extent to which negative environments can help or hinder learners to figure out modal force.

2.2 Modals and negation

Sentences containing modals and negation can in principle receive two interpretations: a 'strong' interpretation (*not* > *possible*, logically equivalent to *necessary* > *not*), and a 'weak' interpretation (*possible* > *not*, logically equivalent to *not* > *necessary*). Cross-linguistically, epistemic possibility modals tend to be interpreted above negation, and root possibility modals below it (Coates 1983, Cinque 1999, Drubig 2001, Hacquard 2010; for a typological overview, see de Haan 1997, van der Auwera 2001). This is illustrated for English in (4a), (4b) and (4c): (root) *can* is interpreted below negation, (epistemic) *might* above negation; *may* is interpreted under negation with a root interpretation, and over negation with an epistemic interpretation.

[11] Other analyses take variable modals to neither be underlying possibility, nor necessity. Kratzer (2012) analyzes them as upper-end degree modals, roughly equivalent in meaning to 'it is somewhat probable (/desirable) that p' (Kratzer 2012).

(4)	a. Jo *can't*$_{\text{root}}$ draw.	¬ ◊	*◊ ¬
	b. Jo *might*$_{\text{epistemic}}$ *not* draw.	*¬ ◊	◊ ¬
	c. Jo *may*$_{\text{root/epistemic}}$ *not* draw.		
	root: 'it is **not possible** that Jo draws'	¬ ◊	*◊ ¬
	epistemic: 'it is **possible** that Jo does **not** draw'	*¬ ◊	◊ ¬

Necessity modals, on the other hand, seem to keep the same scopal behavior with respect to negation, regardless of flavor: they either systematically scope over negation, like *must*/*should* in (5a) (Dutch *moeten*, German *müssen*) (a behavior Iatridou and Zeiilstra 2013 attribute to their being Positive Polarity Items), or under negation, like *need* in (5b) and *have to* in (5c). English *need*, as well as Dutch *hoeven* and German *brauchen*, are commonly analyzed as a Negative Polarity Items (NPI).

(5)	a. Jo *must not/should not* draw.	□ ¬	*¬ □
	epistemic/root: 'it is **necessary** that Jo does **not** draw'		
	b. Jo *needn't* draw.	*□ ¬	¬ □
	epistemic/root: 'it is **not necessary** that Jo draws.'		
	c. Jo *doesn't have to* draw.	*□ ¬	¬ □
	epistemic/root: 'it is **not necessary** that Jo draws.'		

Thus, modals are not uniform in their interaction with negation, neither force-wise nor flavor-wise. This means that for at least some of the modals children have to learn, using negation to infer their force will be problematic. If they expect negation to scope over all modals by default (regardless of force and flavor), cases like (4b) and (4a) will be problematic: (4b) could suggest a necessity meaning for *might* (*need not ~ might not*), and (4a) a possibility meaning for *must* (*can't ~ mustn't*). If learners expect negation to scope over root modals but under epistemic modals (given some more general assumptions about flavor and scope),[12] (4b) is no longer problematic, but (5a) still is. Alternatively, if learners initially assume strong interpretations for any negated modal sentence (following Crain and Thornton's 1998 *Semantic Subset Principle*; see Moscati et al. 2016),[13] cases like (5b), (5c) and (4b) will be problematic. For negation to be helpful in figuring out a modal's force, learners would need to have already figured out how the modal scopes relative to negation, and expect negation to scope differently based on force and flavor. However, it is not clear how they would figure out the right scope relations between modals and negation without knowing the force of the modals.

In the next section, we briefly review findings about children's understanding of modal force and its interaction with negation from the acquisition literature.

[12] See for example Cinque's hierarchy (1999).
[13] Such considerations might explain why necessity but not possibility modals tend to be PPIs.

2.3 Modal force acquisition

Possibility modals like *can* are found early in child productions, by age 2. The literature reports an asymmetry, with root modals appearing earlier than epistemics (Kuczaj and Maratsos 1975; Papafragou 1998; Cournane 2015a,b; van Dooren et al. 2017).[14] Experimental work on children's comprehension usually targets older children (age 4 and up) (Hirst and Weil 1982; Byrnes and Duff 1989; Noveck et al. 1996; Noveck 2001; Ozturk and Papafragou 2015, a.o), and focus on epistemic flavor, using felicity judgment tasks where children have to judge whether a possibility or a necessity statement is true in scenarios where a toy is hidden in one of two boxes. By age 4, children seem to be sensitive to the relative force of modals, when the contrast is made salient by the experimental design, but they still do not behave like adults. First, they tend to over-accept possibility modals when necessity modals are more appropriate (Noveck 2001; Ozturk and Papafragou 2015). This is traditionally blamed on general difficulty with *scalar implicatures* (Barner and Bachrach, 2010; Barner et al. 2011; Chierchia et al. 2001; Skordos and Papafragou, 2014, a.o.): children have trouble accessing the relevant alternatives that the speaker takes for granted, and using them to understand the implicature when asked to judge sentences in isolation. Second, children also tend to accept necessity modals in possibility situations (Ozturk and Papafragou 2015; Koring et al. 2018), a perhaps more surprising result from an adult's perspective: whereas possibility modals are under-informative but logically true in necessity situations, necessity modals are false in possibility situations. Ozturk and Papafragou (2015) argue that children's difficulty with necessity modals stem from (non-linguistic) difficulty reasoning about indeterminate events: in reasoning tasks that introduce indeterminacy, children tend to commit to a possible conclusion before decisive evidence is available, and arbitrarily select one possibility over the other (a tendency sometimes referred to as *premature closure*; Acredolo and Horobin 1987; Bindra et al. 1980, Piéraut-Le Bonniec 1980; Robinson et al. 2006).

A few experimental studies focus on children's interpretation of sentences containing negated modals (Gualmini and Moscati 2009 (*need*); Moscati and Gualmini 2008 (*can*); Moscati and Crain 2014 (Italian *potere* 'can'), Moscati and Gualmini 2009 (Italian *dovere* 'must'), Koring et al. 2018 (Dutch *hoeven* 'need'). Children tend to prefer strong interpretations of negated modal sentences (*not>possible* or *necessary>not*), even when adults prefer weak ones (*possible>not* or *not>necessary*). These studies assume that children already know the underlying force of their modals, and focus on their scope relative to negation. However, children's non adult-like responses could, in principle, be explained by not knowing the force of the modals involved. For instance, one predicts the same responses for Italian *potere non* (where the possibility modal scopes over negation, leading to weak interpretations) if children assume that *potere* expresses possibility and

[14] This asymmetry has been attributed to conceptual and grammatical factors, but it might instead reflect a frequency asymmetry in the input. For how children learn that modals can be used to express various flavors, see van Dooren et al. (2017, *to appear*).

negation scopes over the modal, or if they assume that *potere* expresses necessity and negation scopes under the modal.

We now turn to our studies, which probe more directly the questions of when and how children figure out modal force by investigating their modal input (section 2) and their early productions (section 3).

2 Children's modal input

The goal of this study is to provide an analysis of the modals children are exposed to. We first present quantitative results from a corpus study: how are possibility and necessity modals distributed in actual speech to children? How frequently do they occur with negation? We then present four experiments, using the corpus data, aimed at assessing the informativity of the conversational context as to force. In Experiment 1, based on the *Human Simulation Paradigm* (Gillette et al. 1999), participants have to guess the force of a missing modal in dialogues extracted from the corpus, allowing us to assess the general informativity of conversational contexts depending on force, negation and flavor (epistemic vs. root). Experiment 2 isolates the role of context from possible biases towards possibility or necessity meanings, by showing participants the blanked modal sentence without the dialogue, while Experiment 3 shows the same sentence with all content words replaced by 'Jabberwocky' nonce words. Experiment 4 focuses on a particular feature of the context, namely the desirability of the prejacent as a cue to force for root modals.

2.1 Corpus study

2.1.1 Methods

We used the Manchester Corpus (Theakston et al., 2001) of UK English (CHILDES database, MacWhinney 2000), which consists of 12 child-mother pairs (6 females; age range: 1;09-3;00) recorded in unstructured play sessions. We chose this corpus for its relative density and uniformity of sampling, and early age range. We focused on the period between ages 2;00 and 3;00. All utterances containing modal auxiliaries and semi-auxiliaries (26,598 of 564,625 total utterances; adult: 20,755; child: 5,842; excluding repetitions (6.6%): adult: 19,986; child: 4,844) were coded for force (possibility vs. necessity) (6), presence of negation (7), and flavor (epistemic vs. root) (8).[15] We did not include *will, would, shall* and *going to* as they primarily express future, for which force is a matter of debate (Stalnaker 1978; Cariani and Santorio 2017, a.o.).

(6) Modal lemmas by force:
Possibility: *can, could, might, may; able to*
Necessity: *must, should, need; have to, got to, be supposed to, need to*

[15] We do not differentiate amongst subtypes of root flavors (ability, teleological, deontic…).

(7) Negation:
No negation: 'I **can** go to the pub now.'
Negation:
on main verb: 'I ca**n't** get it'/ 'I must **not** forget Whispy.'
on higher auxiliary: 'we **don't** have to play with your toys.'
on embedding verb: 'I **don't think** you have to look for it.'
other negative quantifier: '**nobody** can reach it.'

(8) Flavor:
Root:

MOTHER:	we won't do that.	
CHILD:	I want her.	
CHILD:	I want her.	
MOTHER:	well you **must** treat her nicely then.	(Aran, 2;07,14)

Epistemic:

MOTHER:	oh.	
MOTHER:	somebody's done a neat pattern, haven't they?	
MOTHER:	goodness me.	
MOTHER:	that **must** have taken a long time.	(Anne, 2;02.10)

2.1.2 Results

We find that overall, possibility modals are more frequent than necessity modals in adult speech: they represent 72.5% of all adults' modal utterances (**Table 1**). This is mostly driven by the highly frequent modal *can*.

Turning to negation and other DE contexts, we find that possibility modals occur with **negation** more frequently than necessity modals (possibility: 20.9%vs. necessity: 10.1%). For necessity modals, negation occurs proportionally more frequently with modals that outscope negation (*should*: 22.8%; *must*: 15.8%, vs. *have-to*: 4.5%; *got-to*: 1.1%). Modals rarely occur with other **negative quantifiers** (*e.g. nothing/never*), with no difference between possibility and necessity (possibility: 0.2%; necessity: 0.1%), nor under a negated embedding verb (*e.g. don't think*), again with no difference between possibility and necessity (possibility: 1.5%; necessity: 2.1%). Details of negative environments are provided in Appendix A (**Table 10**). We further find that modals are extremely rare in **antecedents of conditionals** (0.6% of adults' modal utterances). Necessity modals almost never occur in such environments: we find only 15 occurrences in the whole corpus (106 possibility modals), with 7 of them corresponding to 'if you must'. As a point of comparison, 135 necessity modals occur in consequents of conditionals vs. 432 possibility modals. A breakdown by modal is provided in Appendix A (**Table 11**).

Table 1 Counts and percentages of modal uses by force for adults, ordered by lemma frequency, with and without negation (repetitions excluded: 3.7% of the data).[16]* indicates necessity modals that outscope negation.

	ADULT (n=19,986)		ADULT (n=18,853) [17]			
	all		no negation		negation	
POSSIBILITY	**14,491**	**72.5%**	**10,672**	**79.1%**	**2,828**	**20.9%**
can	11,472	57.4%	8,383	77.7%	2,396	22.2%
could	1,449	7.3%	1,116	96.6%	39	3.3%
might	1,216	6.1%	1,005	82.8%	208	17.1%
able	315	1.6%	134	42.5%	181	57.4%
may	39	0.2%	34	89.5%	4	10.5%
NECESSITY	**5,495**	**27.5%**	**4,814**	**89.9%**	**539**	**10.1%**
have to	2,398	12.0%	2,290	95.5%	108	4.5%
got to	940	4.7%	926	98.8%	11	1.1%
*should**	793	4.0%	537	77.1%	159	22.8%
need (to)[18]	493	2.5%	409	82.9%	84	17.0%
*must**	452	2.3%	346	84.1%	65	15.8%
supposed to	335	1.7%	230	68.6%	105	31.3%
*ought to**	84	0.4%	76	91.5%	7	8.4%

Overall, epistemic uses of modals are rare: they represent only 8.8% of all adults' modal utterances (**Table 2**). Negation is significantly more frequent on root than on epistemic modals (epistemic: 4.6% vs. root: 19.1%). A breakdown by modal is provided in Appendix A (**Table 12**).

Table 2 Counts and percentages of modal uses, by force, flavor and negation, for adults (excluding tags and repetitions)

	ADULT (n=18,853)					
	all		no negation		negation	
root	**17,190**	**91.2%**	**13,896**	**80.9%**	**3,293**	**19.1%**
possibility	12,175	64.6%	9,414	77.3%	2,761	22.6%
necessity	5,015	26.6%	4,482	89.4%	533	10.5%
epistemic	**1,662**	**8.8%**	**1,590**	**95.4%**	**73**	**4.6%**
possibility	1,324	7.0%	1,257	94.9%	67	5.0%

[16] We considered as repetitions cases where the speaker repeated a sentence uttered right before by herself or by another speaker with no significant change.

[17] Excluding tags and repetitions. Tag questions *(e.g.* 'you can wash it later, *can't you*?*'*) are very frequent in this corpus (4.7% of all modal utterances). We decided to exclude modals in tags, as they do not directly matter for our purposes.

[18] There are only 5 occurrences of the NPI *need* (*e.g.* 'you **needn't** whisper.')

necessity	341	1.8%		332	97.3%	6	2.6%

2.1.3 Interim discussion

Overall, possibility modals are more frequent than necessity modals in mother speech: children thus have more opportunities to learn them. The relative rarity of necessity modals may be due to the alternative ways speakers can express necessity (e.g., using imperatives for deontic necessity, or asserting the prejacent directly for epistemic necessity).

Necessity modals rarely appear in downward entailing environments. Negation is infrequent with necessity modals: only 10.1% of all necessity modals cooccur with negation (vs. 20.9% of possibility modals), and necessity modals are exceedingly rare in antecedents of conditionals.

2.2 Experiment 1: adults' modal productions

To assess the general informativity of natural conversational contexts about force, we implemented a variant of the *Human Simulation Paradigm* (Gillette et al. 1999), using dialogue contexts extracted from the corpus. The goal of the original *Human Simulation Paradigm* (Gillette et al. 1998; see also Snedeker 2000; Snedeker et al., 1999; White 2017) is to compare the effect of different kinds of contextual information on the ability to recover a word's meaning: extralinguistic scenes, associated words and morphemes, or syntactic-frame information. The accuracy with which participants can recover the actual word given the context is taken as a general measure of informativity of properties of that context. Following Orita et al. (2013), we use the paradigm in a slightly different way: participants were given only written transcripts from the corpus (with no visual or acoustic information), and had to choose between a possibility and a necessity modal.[19] This allows us to, first, give a general measure of the informativity of conversational context about force: can naïve subjects guess the force of a blanked-out modal based solely on excerpts of conversations in which it appears? Second, we can test directly for interrelationships between force and negation: are contexts equally informative for both necessity and possibility modals? Are negative contexts more informative than positive contexts?

2.2.1 Methods

Procedure. The experiment was run online on IBEX Farm.[20] Participants recruited via Amazon MechanicalTurk were asked to guess a redacted modal in a dialogue between a child and mother by choosing between two options, corresponding either to a possibility (e.g. *might*) or a necessity modal (e.g. *must*), as illustrated in **Figure 1a**. All dialogue contexts consisted of the modal sentence with a blank and the 7 preceding utterances, with the two options displayed at the bottom

[19] Results from another HSP study where participants had to fill in the blank (instead of making a force choice) are reported in Dieuleveut et al. (2019).

[20] An example of the experiment can be accessed at https://farm.pcibex.net/r/HDtxJP/

of the screen. There was first a short training where participants had to choose between the definite vs. indefinite article (*the* vs. *a*) (3 examples with feedback), followed by the test phase without feedback. Overall, each participant had to judge 40 different dialogues (20 trials: 10 possibility, 10 necessity; 20 controls using tense: 10 past, 10 future), presented in random order. The 20 trials were selected randomly for each participant from a list of 40 contexts originally extracted from the corpus; the 20 controls were the same for all participants. Further details of the instructions and material are provided in **Appendix B**.

Conditions. We tested force (possibility vs. necessity) within participants, and flavor (root vs. epistemic) and negation (present vs. absent) between participants. Negation was tested only for root flavor, because negated epistemics were too rare in the corpus (**Table 2**). **Table 3** summarizes the experimental design.

Test condition (between participants)		Modal lemmas	
		possibility	necessity
EPI-AFF (epistemic affirmative)		*might*	*must*
ROOT-AFF (root affirmative)[21]	**ROOT-AFF-1**	*can*	*must*
	ROOT-AFF-2	*can/able*	*have to*
ROOT-NEG (root negative)		*can't/not able*	*not have to*

Table 3. Summary of experimental conditions

Material. ***Extraction procedure*** **–** 160 contexts (2*20 per condition) were randomly extracted from the corpus for the different modals (*can, able*, *might*, *must, have to)*. ***Exclusion criteria*** **–** We excluded contexts where the adult or the child used the target modal in preceding utterances. Contexts were not excluded when the adult or the child used another non-target modal. Briticisms, such as *willn't*, were removed from the dialogue and replaced with American English equivalent (e.g. *won't*). We didn't exclude contexts where there were tag questions (e.g., *'..., mustn't she?'*), but removed the tags when they occurred in the target sentence. ***Controls*** **–** Participants had to choose between future and past (*e.g.* [*saw*] vs. [*will see*], see **Figure 1b**). Importantly, the correct answer was not always guessable based on the target sentence alone: this required participants to read the entire dialogue. Extraction procedure and data cleaning were the same as for targets. We excluded participants that were less than 75% accurate on controls.

[21] We implemented two versions of the **ROOT-AFF** condition. **ROOT-AFF-1** (*can* vs. *must*) allowed us to keep syntactic category of both options identical, while **ROOT-AFF-2** (*can/able to* vs. *have to*) allowed us to avoid concerns related to the formality of *must* for US English speakers. In cases where *have to* was tensed, we used *able to* as the alternative to avoid losing tense information: for example, participants had to choose between [*will have to*] and [*will be able to*]. We extracted the same number of contexts from *able to*, to avoid having the *able to* option always be the wrong answer. Same principles applied for **ROOT-NEG** condition: participants had to choose between [*didn't have to*] and [*wasn't able to*] when *have to* was tensed.

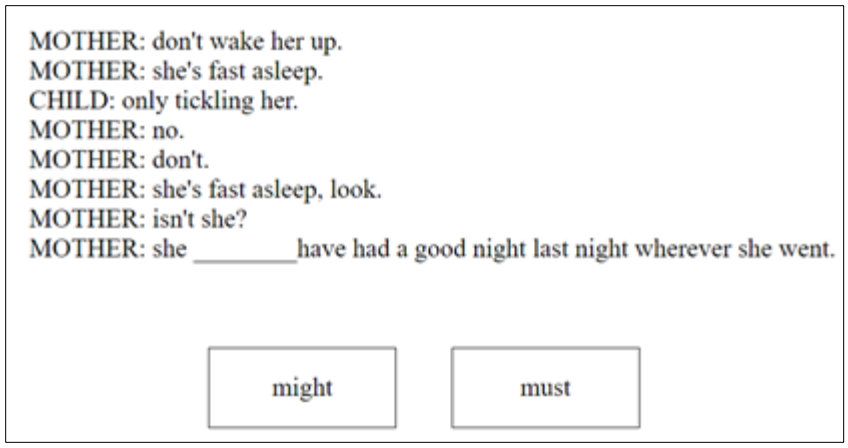


Fig. 1a Experiment 1 stimuli: example trial (EPI-AFF, *must*)

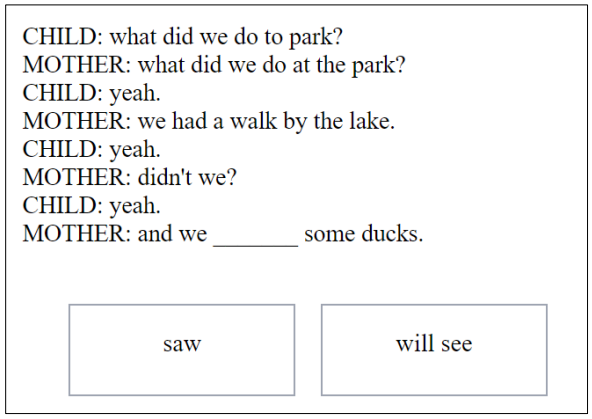


Fig. 1b Experiment 1 stimuli: control trial (*saw*)

2.2.2 Results

Participants. 289 participants were recruited on Amazon Mechanical Turk (4 groups (between participants): ROOT-AFF-1: 73, ROOT-AFF-2: 72; ROOT-NEG: 73; EPI-AFF: 71; language: US English; 156 females, mean age=40.6-years-old). We removed from analysis 8 participants (2.8%) who were less than 75% accurate on controls. We thus present results for 281 participants (ROOT-AFF-1: 71, ROOT-AFF-2: 69; ROOT-NEG: 70; EPI-AFF: 71).

Analysis. All data analyses were conducted using R (R Core Team, 2013), using the package lme4 (Bates *et al*. 2014a, 2014b). All R scripts for analysis are available at https://osf.io/v9ure/. **Figure 2** summarizes the mean accuracy for each condition.[22] Overall, participants were highly accurate at guessing modal force (general mean accuracy: 79.9%). We first ran binomial tests to see whether they differ from chance for each condition (**Table 4**). Participants' accuracy significantly differs from chance in each condition. Their lowest performance is found for ROOT-NEG necessity modals

[22] Accuracy for controls was very high (94.6%). There was no difference between groups in accuracy.

(*e.g. not have to*) (61.3%). **Force –** To test whether there was an effect of Force, we fitted a generalized linear mixed effects model, built with a maximal random effect structure, testing Accuracy (dependent variable, binomial), with Force as a fixed effect and Subject and Item as random factors (following Barr *et al.*, 2013) (glmer syntax: Accuracy~Force+(Force|Subject)+(1|Item),[23] first overall and then subsetting the data for each condition. We compared these models with a reduced model without Force as a fixed effect (Accuracy~1+(Force|Subject)+(1|Item)).[24] We find a general effect of Force, in the direction of a higher accuracy for possibility contexts (Model Comparison: $\chi^2(1)=20.49,\ p=5.9e\text{-}6$***).[25] Restricting to each comparison group, we find a significant effect in ROOT-AFF-1 ($\chi^2(1)=61.1,\ p=5.5e\text{-}15$***) and ROOT-NEG ($\chi^2(1)=15.6,\ p=7.8e\text{-}05$***), again in the direction of a higher accuracy for possibility contexts. In EPI-AFF, the effect of Force almost reaches significance ($\chi^2(1)=3.73,\ p=.053$ *(NS)*). In ROOT-AFF-2, it is not significant ($\chi^2(1)=6e\text{-}04,\ p=0.98$ *(NS)*). **Negation –** We compared ROOT-AFF-2 and ROOT-NEG, as these conditions included the same lemmas. We used the same method as above, comparing a model with Negation as fixed effect and Subject and Item as random factors to a model without Negation as fixed effect (Maximal model: Accuracy~Negation+(1|Subject)+(1|Item); for the interaction: Accuracy~Force*Negation+(1|Subject)+(1|Item)). We find a strong interaction effect (**Interaction Force*Neg**: $\chi^2(1)=7.9,\ p=0.0047$**). We find a significant effect of negation on necessity modals, which leads to lower accuracy (*have to* vs. *not-have to*: $\chi^2(1)=6.5,\ p=0.011$*). On possibility modals, negation leads to higher accuracy, but the effect is not significant (*can* vs. *can't*: $\chi^2(1)=2.29,\ p=0.13$ (NS)). **Flavor –** There was no effect of flavor ($\chi^2(1)=0.11,\ p=0.74$ *(NS)*) (Maximal Model: Accuracy~Flavor+(1|Subject)+(1|Item)).

	Mean accuracy (se) [26]		Exact binomial tests (two-sided)	
	possibility	necessity	possibility	necessity
ROOT-AFF-1	91.7% (0.027)	71.7% (0.054)	p<.001*** 95% CI [0.90, 0.94]	p<.001*** 95% CI [0.68, 0.75]
ROOT-AFF-2	81.5% (0.053)	82.0% (0.052)	p<.001*** 95% CI [0.79, 0.85]	p<.001*** 95% CI [0.79, 0.84]
ROOT-NEG	89.5% (0.031)	61.3% (0.065)	p<.001*** 95% CI [0.88, 0.92]	p<.001*** 95% CI [0.56, 0.64]
EPI-AFF	87.2% (0.028)	74.3% (0.049)	p<.001***	p<.001***

[23] We sometimes had to step back to random-intercepts-only models when the model failed to converge with the full random-effects specification.

[24] We use the same procedure (based on model comparisons) for all subsequent experiments reported in this paper, but won't systematically report the reduced model.

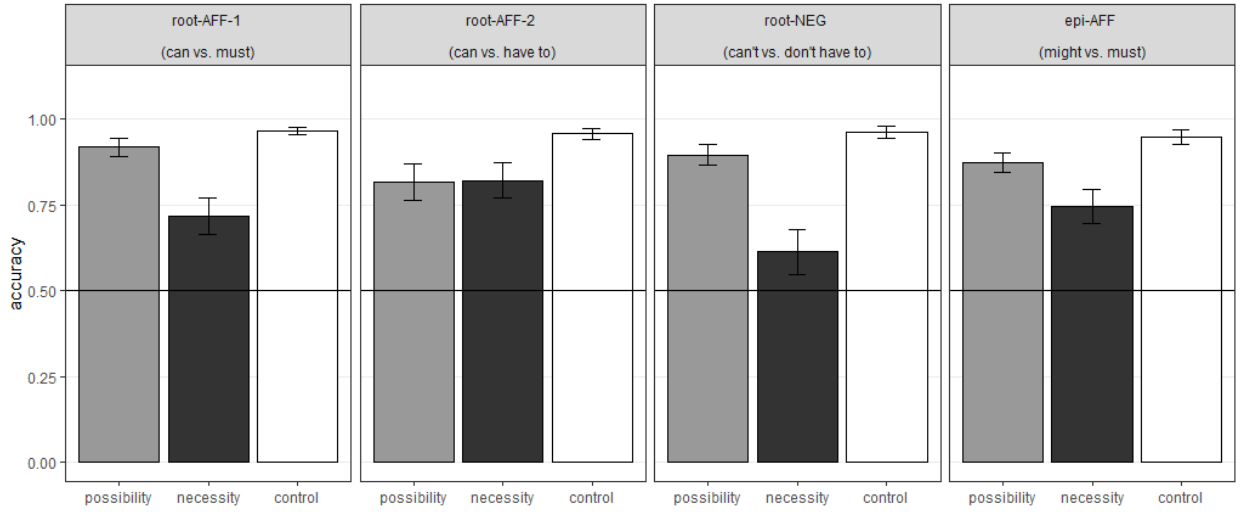
[25] Answers were coded as 1 if the response was accurate, and 0 otherwise. Effect sizes (β), standard errors (SE), z-values, and p-values for the logistic models are reported in Appendix C, for Experiment 1 and all the following analysis.

[26] Accuracy corresponds to the mean accuracy (how good participants were to guess correctly the force of the modal, *e.g.* to answer *can* in a possibility context) across 20 contexts initially extracted from the corpus for each condition of force and flavor. Each participant saw only 10 out of the 20 contexts (10 for possibility, 10 for necessity). On average, each context was thus seen by 34.7 participants (ranging between 24 and 47). See Appendix B, Table 12, for details.

			95% CI [0.84, 0.90]	95% CI [0.71, 0.77]
Total	87.5% (0.018)	72.3% (0.028)		
ALL	79.9% (0.018)			

Table 4 Accuracy rates and binomial tests by condition, testing for significance versus chance (Experiment 1: adults' productions) (n=281, 10 observations per cell)

Fig. 2 Accuracy by condition (adult, n=281*10)



Analysis by contexts (post-hoc). To get a sense of the kinds of contextual cues that were particularly helpful, we examined the contexts that led to lowest and highest accuracy, both for root and epistemic flavors. We focused on necessity modals as there was more variability in accuracy for necessity modals. This informal analysis revealed two factors, depending on flavor: for root modals, cases where the proposition expressed by the prejacent seemed undesirable (e.g., going to the hospital) or effortful (e.g., lifting a heavy object) seemed to lead to high accuracy for necessity modals (see (9)). For epistemic modals, we found high accuracy for necessity modals in contexts that made salient robust evidence for the prejacent (see (10)). Our post-hoc analysis also pointed out a particularly high accuracy for possibility root modals in interrogative sentences (e.g. ___ *you see*?) (mean accuracy for root possibility modals in interrogative: 96.0%).[27] In this case, accuracy may not reflect pure informativity, as participants may rely on idiomatic turns of phrases. However, they were still very accurate restricting to contexts that did not involve interrogatives (mean accuracy for root possibility in declarative: 76.3%).

(9) CHILD: Mummy.
CHILD: Mummy.

[27] Contexts involving interrogative sentences appeared almost exclusively in ROOT-AFF-1 and ROOT-AFF-2, as epistemic and negated modals are rare in interrogatives. Out of 80 contexts for root-AFF, there were 21 interrogative sentences (19 involving possibility modals; 2 necessity modals).

MOTHER: Mummy?
CHILD: that Mummy.
MOTHER: what... what happened to Mummy?
CHILD: poorly.
MOTHER: she's poorly, is she?
MOTHER: she... she _____ go to the hospital.
(*have to,* 'undesirable'; mean accuracy: 96.6%)

(10) MOTHER: ....
CHILD: yeah.
MOTHER: but Bertie was very close behind, wasn't he?
MOTHER: it was a near thing I think.
CHILD: he's lost his hat.
MOTHER: he has.
MOTHER: yeah.
MOTHER: it _____ have been windy eh?
(*must,* 'strong justification'; mean accuracy: 92.1%)

2.2.3 Interim discussion

Results from Experiment 1 show that the conversational context is informative about force: participants were able to guess the force of the modal accurately, just based on short conversation transcripts, and for both forces (general mean accuracy: 79.9%; possibility modals: 87.5%; necessity modals: 72.3%). This means that the information is there, at least in principle, for learners to figure out modal force based on context alone. If children are sensitive to the same cues as adults, they may not need to rely on negation, nor on a bias towards necessity meanings to figure out force.

Multiple factors may play a role in making the conversational context useful for guessing the right modal force: situational cues (e.g., who the interlocutors are), cues from world knowledge (e.g., what is allowed or prohibited), or pragmatic cues (what the speaker is trying to achieve, in particular performing orders, permissions or prohibitions). Our post-hoc exploration suggests that the cues may vary based on modal flavor. It appears that the (un)desirability and effortfulness of the prejacent could be particularly useful for roots, and some explicit justification for epistemics. We probe the effect of desirability more directly in Experiment 3 below.

Of course, some of the cues available to adults in this experiment might not to be usable by children: for instance, children might lack some world knowledge. This limitation is intrinsic to any paradigm where adults are used to simulate word learning (the task asked of the adults is to guess a word they already know, whereas children have to guess the meaning of a *new* word from the context in which it is used) (see White 2017, Orita 2013, for discussion). That said, children

also have access to a substantially richer context than participants in our experiment, who had no visual nor prosodic information, and no common ground with the child and the mother.

We find a general effect of force, with participants being more accurate with possibility modals. This could be interpreted as possibility contexts being more informative than necessity contexts. However, the effect is carried by only 2 sub-conditions (ROOT-AFF-1 and ROOT-NEG; it is near-significant in EPI-AFF ($\chi^2(1)=3.73$, $p=.053$), and not significant in ROOT-AFF-2). It is not significant once we take into account the effect of interrogative sentences, which lead to a very high accuracy for root possibility modals (if we restrict to declarative contexts only, participants don't perform significantly better on possibility contexts).[28]

Lastly, turning to negation, we find that negated necessity modals are rare: our corpus results show that overall, modals scoping under negation are negated 7.4% of the time, (vs.22.6% for root possibility modals).[29] The results from Experiment 1 show that they are also less informative. We find opposite effects of negation on possibility and necessity modals: while negation leads to a slightly higher accuracy for possibility modals (*can't*: 89.5% vs. *can*: 81.5% (*NS*)), it leads to lower accuracy for necessity modals (*don't have to*: 61.3% vs. *have to*: 82.0%, $p=0.011$*) (significant interaction effect Force*Negation: $p=0.0047$**).

Why is that so? First, the low frequency of negated necessity modals may come from a competition with the use of a bare possibility modal, which can convey non-necessity via a scalar implicature (Horn 1972).[30] We find a few cases that could be informative for children, like (11), where the context makes clear that the impossibility interpretation does not hold.

(11) MOTHER: now we don't throw things, do we?
MOTHER: now.
CHILD: I don't want that anymore.
MOTHER: well then.
MOTHER: we won't play with it anymore.

[28] This higher accuracy in possibility contexts might also reflect a general tendency to answer with possibility modals by default, maybe because of their relative frequency. To test for the effect of frequency, we compared accuracy for *can* and *able-to* (used in ROOT-AFF-2 and ROOT-NEG), which are both root possibility modals but strongly differ in frequency (3 *able* for 100 *can* in the Manchester corpus). The general accuracy on *able* was not significantly lower than on *can* (overall: *able*: 80.8% vs. *can*: 89.8%; vs. *have to*: 71.7%).

[29] Few other corpus studies address the distribution of possibility and necessity modals with respect to negation, and fewer look at child-directed speech, but they also suggest that negated necessity modals are not frequent. De Haan (2011) reports that negation is very rare with *must*: 2.5% in the Brown corpus (written English), and 1.4% in the Switchboard corpus (spoken English). Thornton and Tesan (2013) report the frequencies of some negative auxiliary verbs in the input to children in the Providence corpus, but don't specify their frequency relative to the positive forms. Jeretič (2019) reports that negation on necessity modals is infrequent in the input to French and Spanish children (French: 15.5%;Spanish: 6.2%).

[30] Horn focuses on a different but related problem, namely the fact that cross-linguistically, the 'O' corner of the Aristotelian square of opposition (corresponding to negated universals, e.g., non-necessity meanings) seems to never be lexicalized, whereas the other three corners (possibility, necessity and impossibility) can be. Horn argues that this follows from the fact that there is no functional pressure to lexicalize non-necessity meanings: speakers already have a way to express non necessity, using s*calar implicatures*.

MOTHER: you **don't have to** play with it.
MOTHER: you can play with something else. (Aran, 2;6.10)

However our results suggest that most adult negated necessity modals are cases like (12), where the conveyed meaning is close to impossibility, which illustrate 'polite' uses of negated necessity modals. Here, *don't have to* is used to perform a prohibition.

(12) CHILD: break.
MOTHER: you want me to break it?
CHILD: yeah.
MOTHER: no.
MOTHER: we **don't have to** break these things.
MOTHER: oh.
MOTHER: you've broken it.
CHILD: yeah. (Aran, 2;0.28)

From this, we conclude that it is unlikely that children rely on negation to figure out the force of necessity modals. First, as discussed earlier, negation is potentially misleading for a number of necessity modals: *mustn't* is truth-conditionally equivalent to *can't*, which might drive children to infer that they express possibility, if children assume that negation scopes over root modals by default. Second, necessity modals that can scope under negation (e.g. *have to, got to*) are rare in the input, and their use is particularly misleading about their force because they often can be used to convey impossibility. Children will therefore need other strategies to solve the subset problem. However, our findings suggest that negation could be more helpful to figure out the force of possibility modals: they cooccur frequently in the input (22.6% of root possibility modals are negated), and Experiment 1 shows that impossibility contexts are highly informative (mean accuracy for *can't*: 89.5%). Children may be able to infer from these occurrences the force of possibility modals, if they expect negation to scope over modals.

2.3 Experiment 2 and 3: Isolating the role of context

Experiment 1 shows high accuracy for both possibility and necessity. We take these results to mean that the context is informative as to force. But could it be that participants succeed at the task not by relying on the context, but through biases, which could also be at play in children's modal learning? In particular, could their high accuracy be due to a necessity bias that allows them to correctly guess necessity meanings?[31] To isolate the contribution of the dialogue context, we ran two related follow-up experiments, presenting either only the target sentence without its discourse context (Experiment 2, 'sentence-only') or presenting only the target sentence but replacing all content words with nonce words (Experiment 3, 'Jabberwocky'). We expect that participants'

[31] We thank an anonymous reviewer for pointing this issue out.

performance should decrease in Experiments 2 and 3, if their successes in Experiment 1 was due to a reliance on context.

2.3.1 Methods

Procedure. Experiment 2 and 3 were identical to Experiment 1, except that in Experiment 2, participants only saw the target sentence (without the preceding dialogue, see **Figure 3a)** and in Experiment 3, they saw only the target sentence with all content words replaced by Jabberwocky (see **Figure 3b**).[32] Experiment 2 lets us isolate the specific contribution of the dialogue. However, it does not remove all contextual information, since the content of the prejacent contributes to context (e.g., desirability of the event it describes). This motivates Experiment 3, which further removes any semantic information in the sentence (see Gillette 1998, White et al. 2017, a.o.). As the task was shorter, participants judged all 40 contexts (Experiment 2: 60 trials: 20 possibility; 20 necessity; 20 controls using tense; Experiment 3: 40 trials: 20 possibility; 20 necessity). To make sure that participants kept paying attention, we also had 8 attention checks (simple additions and subtractions, *e.g.* 1+3=_____). We removed from target sentences any repetitions (*e.g.* '*dolly... dolly _____ use her pottie'* was corrected to '*dolly _____ use her pottie'*), as well as phatic words (*e.g. oh, yeah*). For Experiment 3, we replaced all content words (nouns, verbs, adjectives, adverbs) with nonce Jabberwocky words (*e.g. shink, gumbly*). We kept all function words: determiners, prepositions, complementizers, connectives, personal pronouns, temporal adverbs and locatives, auxiliaries (*be*, *have*, modals, semi-modals, other than the target modal, which was replaced by a ____), and all plural morphology, and tense and aspect marking). Conditions were the same as in Experiment 1. Instructions are provided in **Appendix B**.

Fig. 3a Experiment 2 stimuli (EPI-AFF, *must*)

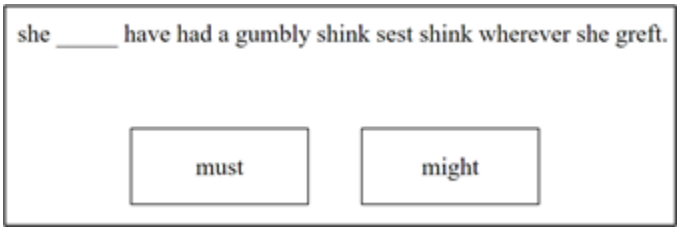
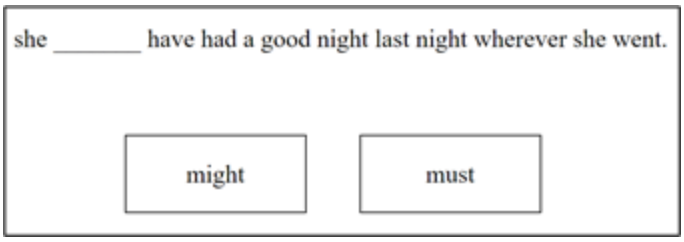


Fig. 3b Experiment 3 stimuli (EPI-AFF, *must*)

[32] Examples of the experiments can be accessed at
2: https://farm.pcibex.net/r/sohaoF/
3: https://farm.pcibex.net/r/ZbVcQT/



2.3.2 Results

Participants. 252 participants were recruited on Amazon Mechanical Turk (Experiment 2: ROOT-AFF-1: 31, ROOT-AFF-2: 33; ROOT-NEG: 30; EPI-AFF: 29; 66 females, mean age: 44.0-years-old; Experiment 3: ROOT-AFF-1: 31, ROOT-AFF-2: 29; ROOT-NEG: 30; EPI-AFF: 38; 67 females, mean age: 38.8-years-old). In Experiment 2, we removed from the analysis 1 participant who was less than 75% accurate on attention checks and 6 participants who were less than 75% accurate on tense controls (5.7%).[33] In Experiment 3, accuracy on attention checks was very high (99.1%). No participant was excluded.

Analysis. Table 5a and **5b** report mean accuracy in each condition for Experiment 2 and 3. As for Experiment 1, we first ran binomial tests to see whether they differ from chance for each condition. In Experiment 2, participants were overall still good at guessing force (**Table 5a**). In Experiment 3, participants are above chance for possibility modals, but don't differ from chance for necessity modals except in EPI-AFF (**Table 5b**). We use generalized linear mixed effect models to test the effect of Experiment on Accuracy, first overall and then subsetting the data to test by each subcomparison group, comparing Experiment 2 and 1, 3 and 2, and 1 and 3. We test Accuracy (dependent variable), with Experiment as a fixed effect and Subject and Item as random factors (Maximal model: Accuracy~Experiment+(1|Subject)+(1|tem)). We compared these models with a reduced model without Experiment as a fixed effect (Accuracy~1+(Force|Subject)+(1|Item)). First comparing Experiment 1 and 2 , we find that participants performance is overall lower without the dialogue ($\chi^2(1)=48.2$, $p=3.9e\text{-}12$***). Looking at the 8 subcomparision groups, we find decreased performance for necessity contexts in ROOT-AFF-1, ROOT-AFF-2 and EPI-AFF, and for posibility contexts in ROOT-AFF-2 and ROOT-NEG. We find no difference for possibility ROOT-AFF-1 and EPI-AFF and necessity ROOT-NEG. Detailed results of the comparisons are given in **Appendix D**. Comparing Experiment 2 and 3, we find a significantly lower accuracy for Experiment 3 overall ($\chi^2(1)=188$, $p<2.2e\text{-}16$***) and looking at each subcomparison group. Last, comparing Experiments 1 and 3, we also find decreased performance for Experiment 3 overall ($\chi^2(1)=650$, $p<2.2e\text{-}16$***) and in each subcomparison group (see **Appendix D**. **Figure 4** summarizes the comparison between the three experiments for all conditions.

[33] Accuracy on attention checks and tense controls was very high (attention checks: 99.4%; tense controls: 95.8%), with no difference between groups. To compute accuracy on tense controls, we only included sentences that could not lead to an ambiguity (e.g. because of containing a temporal adverb) (10 out of 20 cases).

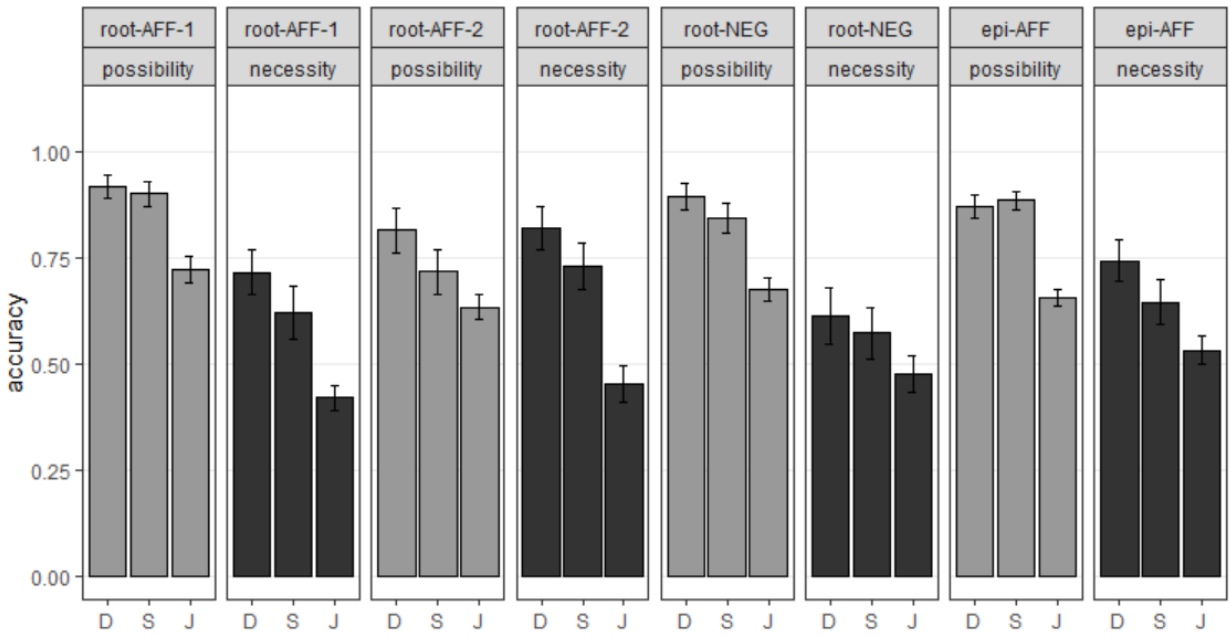
	Mean accuracy (se)		Exact binomial tests (two-sided)	
	possibility	necessity	possibility	necessity
ROOT-AFF-1	90.2% (0.030)	62.0% (0.062)	p<.001*** 95% CI [0.88, 0.92]	p<.001*** 95% CI [0.59, 0.65]
ROOT-AFF-2	71.8% (0.052)	73.0% (0.054)	p<.001*** 95% CI [0.68, 0.74]	p<.001*** 95% CI [0.70, 0.76]
ROOT-NEG	84.8% (0.036)	57.3% (0.061)	p<.001*** 95% CI [0.82, 0.87]	p=.00019*** 95% CI [0.54, 0.61]
EPI-AFF	88.6% (0.021)	64.6% (0.054)	p<.001*** 95% CI [0.86, 0.90]	p<.001*** 95% CI [0.61, 0.68]

Table 5a Accuracy rates and binomial tests by condition, testing for significance versus chance (Experiment 2) (n=116, 20 observations per cell)

	Mean accuracy (se)		Exact binomial tests (two-sided)	
	possibility	necessity	possibility	necessity
ROOT-AFF-1	72.2% (0.031)	42.0% (0.029)	p<.001*** 95% CI [0.69; 0.75]	p=1 (NS) 95% CI [0.39; 0.45]
ROOT-AFF-2	63.4% (0.029)	45.3% (0.043)	p<.001*** 95% CI [0.60;0.67]	p=0.99 (NS) 95% CI [0.42; 0.48]
ROOT-NEG	67.7% (0.027)	47.7% (0.041)	p<.001*** 95% CI [0.64; 0.71]	p=0.88 (NS) 95% CI [0.44; 0.52]
EPI-AFF	65.8% (0.019)	53.3% (0.032)	p<.001*** 95% CI [0.63; 0.69]	p=0.038* 95% CI [0.50; 0.56]

Table 5b Accuracy rates and binomial tests by condition, testing significance versus chance (Experiment 3) (n=129, 20 observations per cell)

Figure 4 Accuracy by condition comparing Experiment 1 (D: 'dialogue'), Experiment 2 (S: 'sentence-only') and Experiment 3 (J: 'Jabberwocky').



2.3.3 Discussion

These two control experiments let us isolate the specific contribution of context, and show that it is informative beyond potential biases. In the Jabberwocky version (Experiment 3), participants' mean accuracy is 57.2%, vs. 74.0% when they see the sentence with content words (Experiment 2), and 79.9% when they see the preceding dialogue (Experiment 1). This first confirms that syntax alone doesn't help: when we remove all semantic information, subjects are at chance.[34] Moreover, further analysis looking at the interaction between Force and Experiment (reported in Appendix D, Table Y) suggests that the effect of having more contextual information is stronger for necessity modals than for possibility modals. If participants' high accuracy in Experiment 1 was due to a necessity bias, we would expect their performance to remain the same in Experiment 2: participants should guess necessity meanings, unless presented with direct evidence against it. Altogether, participants' high accuracy on possibility modals, even in the absence of context, suggests that if they bring a force bias to the task, it is more likely to be a possibility, rather than a necessity bias.

2.4 Experiment 4: desirability

The results from experiments 1-3 argue that the conversational context in which modals are used is informative about their force. But what is it about the context that is particularly informative?

[34] The fact that participants perform better than chance on possibility contexts in the Jabberwocky experiment might reflect a general tendency to answer with possibility modals when given a force choice task—perhaps because of their higher frequency, see footnote 29). It may also be driven by their higher accuracy in interrogative sentences, as discussed in section 2.2.3 (mean accuracy for possibility in interrogatives: Exp1: 98.1%; Exp2: 96.2%; Exp3: 76.1%; in declaratives: Exp1: 75.1%; Exp2: 65.8%; Exp3: 59.5%).

As discussed in section 2.2.3, several factors could be at play. Our post-hoc analysis suggested that the cues may vary with flavor: for root modals, necessity modals seem associated with undesirable and effortful events; for epistemics, necessity modals seem to occur in contexts that highlight strong evidence that support the proposition expressed by the prejacent. We now turn to an experiment that tests the hypothesis that (un)desirability matters for root modals, as an initial proof of concept, and leave a more systematic probing of additional features of the context for future research.

We hypothesize that the desirability of the prejacent could be playing a crucial role in the acquisition of force for root modals. Desirabilityis a feature likely to be conceptually accessible to young children: the cognitive developmental literature suggests that children can reason about desires quite early on, and understand that people can have incompatible desires (Wellman and Woolley 1990, Repacholi and Gopnik 1997, Rakoczy et al. 2007, Ruffman et al. 2017, a.o.). Moreover, preschool children have been shown to be sensitive to desirability for modal usage pragmatics, in particular compared to unmodalized expressions (Ozturk and Papafragou 2015). The first goal of Experiment 3 is to assess the availability of this cue in the input: do adults actually use necessity modals more frequently with undesirable events (e.g., 'You *must/#can* eat your brussels sprouts'), and possibility modals with desirable events (e.g., 'You *can/#must* have a cookie')? Second, does this contribute to participants' performance in Experiment 1, i.e., did adults actually rely on this cue to guess force?

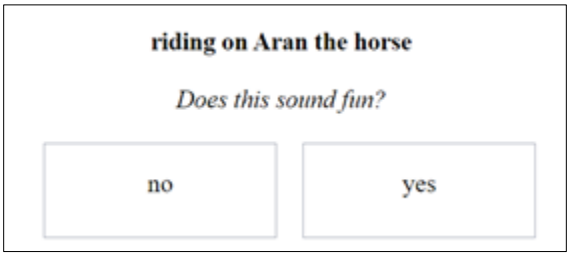
2.4.1 Methods

Procedure. Participants were asked to indicate whether various activities (*e.g.* 'doing a puzzle') sounded fun or not (see **Figure 5**). They were told that the activities involved two-year-old children and their mothers. The different activities corresponded to the prejacents[35] of the modals tested in Experiment 1:[36] for example, for 'Can the dolly ride on Aran the horse?', participants were asked whether '*riding on Aran the horse*' sounded fun ('yes') or not ('no'). We used the prejacents, rather than the full modal sentences to avoid biasing towards positive responses for possibility modals, and negative responses for necessity modals. Referential pronouns (e.g. *it*) were replaced whenever they could be recovered from the context (*e.g.* '*Finding the green marker'* for 'Can you find it?'). In each group, participants judged all 40 prejacents (42 trials: 20 possibility, 20 necessity; 2 initial practice items, which were removed from the analysis). To make sure participants kept paying attention, we had 10 attention checks (*e.g.* 1+3=____). Instructions are given in **Appendix B**. As our hypothesis concerns root modals, we ran the experiment only on ROOT-AFF-1 (*can* vs. *must*) and ROOT-AFF-2 (*can/able* vs. *have to*). ***Rationale***. This experiment allows us to first assess the desirability of the different events in as objective a way as possible, to see if there is a relation between desirability (measured by the proportion of yes answers to 'being fun', a child-friendly

[35] This is not true *stricto sensu*, as participants also lose the information about the subject (e.g. *I/you/Caroline*…).
[36] An example of the experiment can be accessed at: https://farm.pcibex.net/r/lrxZaB/

way of assessing what is desirable) and force usage in the corpus. We can then probe whether adults used this cue to infer force in Experiment 1 by looking at the correlation between the desirability score in Experiment 3 and accuracy in Experiment 1. We expect a negative correlation for necessity modals (fewer 'yes' responses for accurate guesses of necessity uses) and a positive one for possibility modals (more 'yes' responses for accurate guesses of possibility uses).

Fig. 5 Experiment 4 stimuli: example trial (ROOT-AFF-1, *can*)



2.4.2 Results

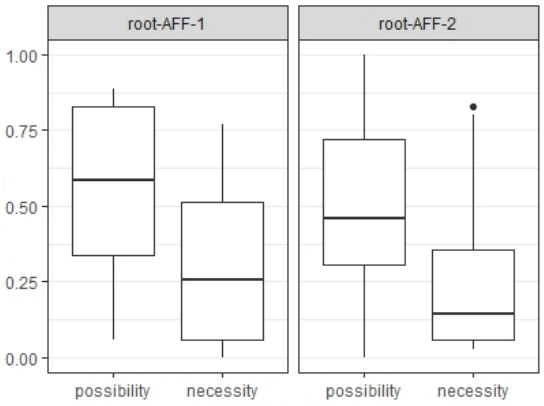
Participants. We recruited 70 participants on Amazon Mechanical Turk (ROOT-AFF-1: 35, ROOT-AFF-2: 35; language: US English; 35 females, mean age: 40.4-years-old). Accuracy on attention checks was very high (99.6%), and we did not have to remove any participant from the analysis based on attention checks.

Analysis. Table 6 reports means of 'yes' answers ('fun') in all conditions. First, to test for the effect of Force (possibility vs necessity), we fitted a generalized linear mixed effects model, testing Answer (dependent variable), with Force as a fixed effect and Subject and Item as random factors. We compared this model with a model without Force as fixed effect (glmer syntax for the maximal model: (Answer)~Force+(1|Subject)+(1|Item); for the reduced model: ((Answer)~1+(1|Subject)+(1|Item)). 'Fun' answers were coded as '1'. We find a general effect of Force: participants judged prejacents extracted from possibility statements overall more 'desirable' than those extracted from necessity statements (Model comparison: $\chi^2(1)=15.5$, $p=8.2e\text{-}05$***). We subsetted the data for each group, using the same formulae and comparison methods. The effect is significant in both groups (ROOT-AFF-1: $\chi^2(1)=8.2$, $p=0.0041$** ; ROOT-AFF-2: $\chi^2(1)=6.2$, $p=0.012$*).[37] **Figure 6** shows the distribution of ratings for possibility and necessity for the two groups. Then, we computed correlations between the desirability score (Experiment 4) and accuracy in Experiment 1 (see **Figure 7**). For possibility, we find a weak positive correlation (Pearson's $r=0.12$) ($t(1398)=4.42$, $p<0.001$**; 95%-CI: [0.065; 0.168]); for necessity, a weak

[37] We sometimes had to step back to random-intercepts-only models when the model failed to converge with the full random-effects specification. We also checked that there was no significant difference between groups (ROOT-AFF-1 vs ROOT-AFF-2) ((Answer=1)~Group+(1|Subject)+(1|Item). Results of the model comparison were not significant ($\chi^2(1)=0.22$, $p=0.64$).

negative correlation (Pearson's *r=-0.073*) (t(1398)=-2.74*, p=0.0063****; 95%-CI: [-0.125; -0.021]).

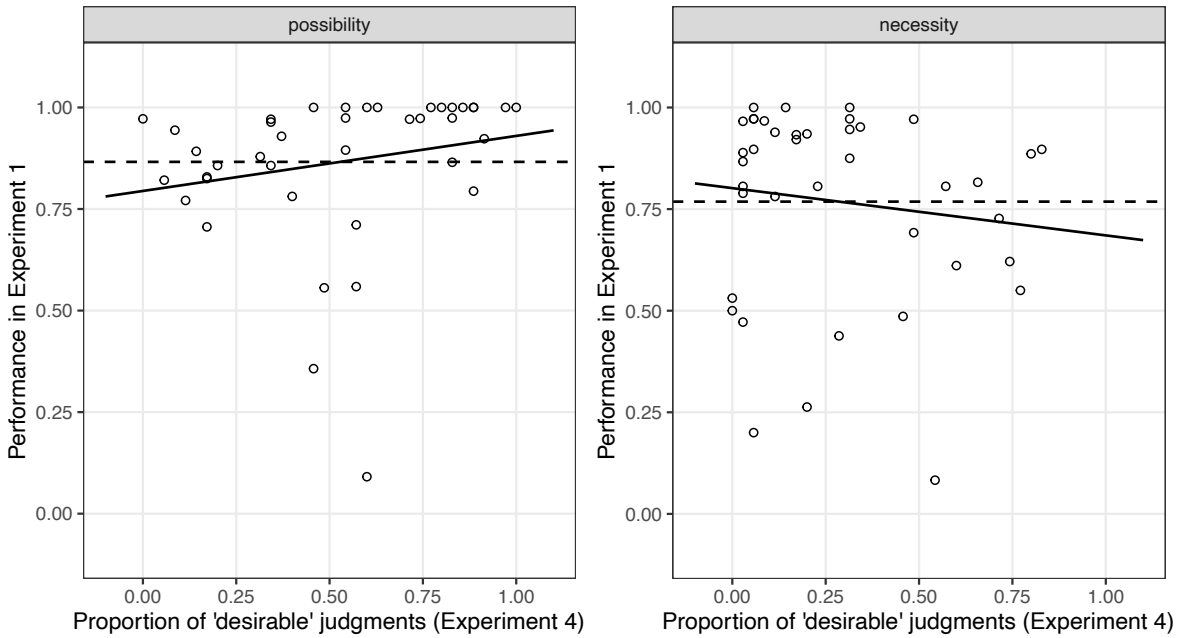
Fig. 6 Distribution of 'desirable' answers for possibility and necessity contexts for each group



	Mean of desirable ('yes') answers (se) possibility	necessity	Effect of Force
ROOT-AFF-1	56.0% (0.063)	31.4% (0.060)	*χ^2(1=8.22, p=0.0041***
ROOT-AFF-2	49.7% (0.067)	25.7% (0.057)	*χ^2(1)=6.2562, p=0.012**
ALL	52.9% (0.045)	28.6% (0.041)	*χ^2(1)=15.5, p=8.2e-05****
ALL	40.7%		

Table 6 Desirablity scores and significance tests (binomial linear mixed effects models comparing possibility/necessity) for possibility and necessity modals

Fig. 7 Relation between accuracy in Experiment 1 (y-axis) and desirability score in Experiment 4 (x-axis) by force. Black lines correspond to Pearson's r. Dashed lines correspond to the mean accuracy in Experiment 1, for possibility and necessity contexts.



2.4.3 Discussion

Our results confirm our initial observations for Experiment 1, and show that there is a relation in children's input between the desirability of the prejacent (evaluated by participants that were blind to the force of the modal originally used) and force. Adults use possibility modals more frequently with desirable events, and necessity modals with undesirable events (mean desirability score for possibility modals (*can/able*): 52.9%; for necessity modals (*must/have to*): 28.6%). Furthermore, the lower accuracy for possibility modals with undesirable prejacents in Experiment 1 and for necessity modals with desirable prejacents suggests that adult participants in Experiment 1 likely made use of desirability in their force judgments. Together, this suggests that children could conceivably use this cue: it is available in the input, and the cognitive developmental literature suggests they are sensitive to desirability, though its association with obligations may require some life experience.

2.5 Summary: children's modal input

Our corpus results show that children are exposed to much more possibility modals than necessity modals, and that they hear the former relatively more often with negation than the latter. We have also seen that negation is unlikely to help learners figure out necessity modals, and might in fact hinder their acquisition. It appears however to be potentially much more helpful for possibility modals. If learners can't rely on negation or other downward entailing environments to solve the subset problem, do they need to rely on a necessity bias? Our experiments suggest that they may not need to, as the conversational context in which modals are used is informative about both forces. If children are able to make use of these conversational cues, they neither need to rely on negation, nor on a necessity bias. One aspect of the context that could be particularly helpful for

root modals is the desirability of the prejacent. Now that we have a clearer picture of children's modal input, and what, in principle, learners may be able to rely on or not, we turn to our study of children's productions.

3. Children's modal productions

To study children's modal production, we used the same methods as for the input study. We first present the results from our corpus analysis, comparing children's early productions to those of adults', and then present results from Experiment 5, which is based on the same paradigm as Experiment 1, but tests children's utterances.

3.1 Corpus study

3.1.1 Results

Like adults, children produce more possibility modals than necessity modals, and the difference is even stronger (79.3% of children's modal productions, vs. adults: 72.5%) (**Table 7**). *Can* is by far their most common modal (75.6%, vs. adults: 57.3%), and *have to* their most frequent necessity modal (7.3%, vs. adults: 12.0%). Necessity modals are particularly rare with negation in their productions (only 5.1%), whereas negated possibility modals are very frequent: 51.0% (adults: necessity modals with negation: 10.1%; possibility modals with negation: 20.9%). Epistemic modals are overall very rare in child productions: they represent only 2.4% of children's modal utterances (114 cases, possibility: 93, necessity: 21) (vs. 8.8% of all adults' modal utterances). Looking at the evolution of children's productions during the time period, summarized in **Figure 8a**, we find that the relative proportion of necessity modals tend to increase with age: while they represent 12% of children's modal productions between 2 and 2;03-year-old, they represent 24.5% between 2.9 and 3-year-old (**Figure 8b** confirms that for adults, the relative proportion of possibility and necessity modals does not significantly change over time: we only find a slight increase of necessity modals).

Table 7 Counts and percentages of modal uses by force, ordered by lemma frequency, with and without negation, for children (repetitions excluded: 17.0% of the data) (*X2 (1, N=24830)=92.6, p<2.2e-16*) [38]

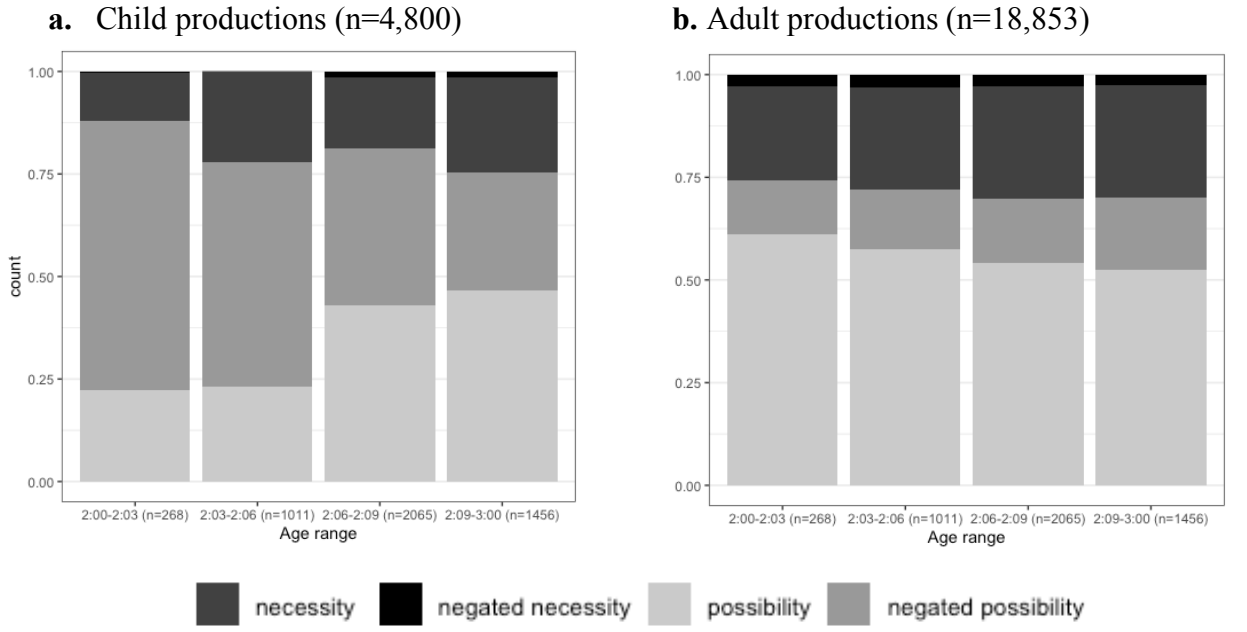
CHILD (n=4844)	CHILD (n=4800) [39]	
all	no negation	negation

[38] Note that the chi-square assumption of independence of observations is violated by corpus samples, as the same speaker supplies multiple uses per cell. However, this test metric is commonly used in corpus linguistics for simple distributional comparisons, and is not straightforwardly a violation as we are comparing spontaneous utterances, not individuals (each spontaneous production is taken as a proxy for independence).

[39] Excluding tag questions (0.9% of children's modal productions, vs. 5.7% of adults' modal productions).

POSSIBILITY	3841	79.3%		1861	49.0%	1937	51.0%
can	3663	75.6%		1739	48.0%	1881	51.9%
might	86	1.8%		78	97.5%	2	2.5%
could	80	1.6%		34	39.5%	52	60.4%
able	3	0.1%		1	33.3%	2	66.6%
may	9	0.2%		9	100%	0	0%
NECESSITY	1003	20.7%		950	94.8%	52	5.2%
have to	352	7.3%		345	98.0%	7	1.9%
got to	288	5.9%		283	98.3%	5	1.7%
should	22	0.5%		17	80.9%	4	19.0%
need to	217	4.5%		204	94.0%	13	5.9%
must	114	2.4%		94	82.5%	20	17.5%
supposed	9	0.2%		6	66.7%	3	33.3%
ought to	1	0.0%		1	100%	0	0%

Fig. 8 Evolution of children's modal productions from 2 to 3-years-old by force and negation, binned in 3 months period



3.1.2 Discussion

We find that children use (root) possibility modals frequently, both with and without negation, which we can take as initial evidence of productivity (Stromswold 1990). Children produce few

necessity modals, and rarely with negation.[40] This difference might be explained by several factors: the difference in frequency in their input (if children grasp more frequent words first), and the differences in social status and topics of conversations between children and adults (for instance, children may be less prone to giving orders and thus less prone to using root necessity modals). Necessity modals tend to become more frequent over time. However, quantitative production data can only provide a partial picture of whether children use and understand modals correctly. To assess these productions in a more qualitative way, we ran Experiment 5 (identical in method to Experiment 1) on children's modals.

3.2 Experiment 5: children's modal productions

The goal of this experiment is to investigate children's early modal productions to see whether they use modals in an adult-like way, by comparing their usage to adult usage (Experiment 1). Can adults guess the force of a modal used by a child, given the conversational context in which they use it?

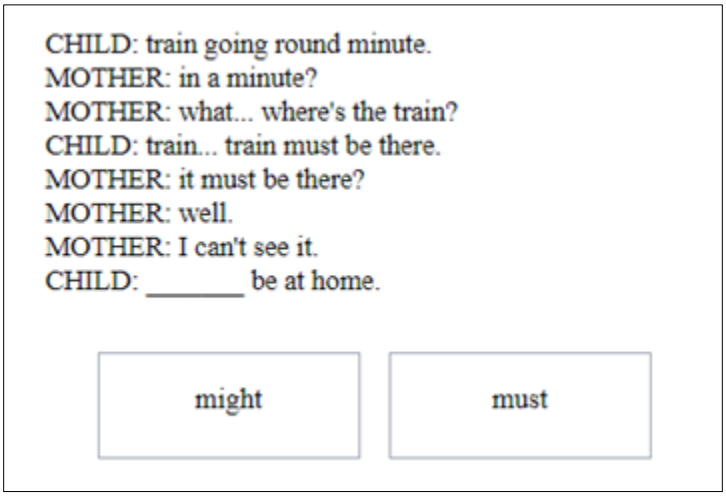
3.2.1 Methods

Experiment 5 was identical to Experiment 1, except that we used children's utterances instead of adults' and made small changes in the instructions (see **Appendix B**).[41] An example of the display is given in **Figure 9**. We implemented the same conditions: ROOT-AFF-1; ROOT-AFF-2; ROOT-NEG; EPI-AFF. Controls were also based on tense (past vs. future). ***Extraction procedure*** – Given the low frequency of negated necessity modals and epistemic necessity modals in child productions, we could test only 10 different contexts for ROOT-NEG necessity and 12 contexts for EPI-AFF necessity conditions.[42] This did not make a difference for the participants, who always had 10 contexts to judge per condition (40 dialogues in the whole experiment: 20 trials: 10 possibility, 10 necessity; 20 controls: 10 past, 10 future). In all the other conditions, the 10 contexts were selected randomly out of a list of 20 contexts initially extracted from the corpus, in the same way as for the adult experiment. ***Exclusion criteria*** – Given the low frequency of negated necessity and epistemic modals, we didn't remove cases where the modal already appeared in the preceding dialogue.[43] We made sure to include examples in the training (*the/a*) and control items (past/future) where the right (or wrong) answer appeared in the preceding dialogue. Again, we removed Briticisms, but we did not correct children's ungrammatical utterances (e.g. *comed* for *came*), except in the case of *have to* when children omitted *to* (so participants would not reject the answer because of its ungrammaticality). ***Rationale*** – We make the assumption that adults rely on their own competence

[40] Similar distributional patterns (possibility modals used more frequently than necessity modals and occurring frequently with negation) are found in Spanish and French (Jeretić 2018) and Dutch (van Dooren et al., in prog.)
[41] An example of the experiment can be accessed below: https://farm.pcibex.net/r/mYOcOz/
[42] Because some of the negated *have to* in child productions were particularly unclear (e.g. 'I *can't have to* read it.'), we also used *not gotta* and *not need*. Details are provided in Appendix B.
[43] Out of 148 contexts, 36 of them had the modal appear in the preceding dialogue (24.3%) (uttered by the child: 11, by the mother or another adult: 20; by both: 5).

to judge usage, and that the dialogues preceding the modal sentence are equally informative for adults' and children's utterances.[44] If children use their modals in an adult-like way, we expect no difference in accuracy between Experiment 5 and Experiment 1. If they do not (*e.g.* they use *can* in a necessity situation, when adults would use *must*, or they use *must* in a possibility situation, when adults would use *can*), we expect a lower accuracy for children's utterances.

Fig. 9 Experiment 5 stimuli (child productions): example trial (*must*)



3.2.2 Results

Participants. 289 participants were recruited on Amazon Mechanical Turk (EPI-AFF: 74, ROOT-AFF-1: 72, ROOT-AFF-2: 73; ROOT-NEG: 72; language: US English; 173 females, mean age=40.2-years-old). We removed 18 participants (6.2%) who were less than 75% accurate on controls.[45] We thus present results for 273 participants (EPI-AFF: 68; ROOT-AFF-1: 70; ROOT-AFF-2: 70; ROOT-NEG: 65).

Analysis. Table 8 reports mean accuracy in each condition (summarized in **Figure 10**). We first ran the same binomial tests as for Experiment 1. Participants performed better than chance in all conditions involving possibility, but not necessity: for ROOT-AFF-2 (*have to*) (mean accuracy: 42.6%) and ROOT-NEG necessity (*not have to*) (mean accuracy: 32.3%), they performed lower than chance (**Table 8**). **Force** — We used generalized linear mixed effects models testing Accuracy

[44] As a proxy, we checked that the mean length of dialogues was the same for adults and children (mean number of words for children's' contexts: 39.6 words/dialogue; for adults' contexts: 38.9).
[45] For the adult version, the proportion of errors on controls was very low (5.4%), with no difference between groups. For the child version however, the initial proportion of errors on controls was quite high (21.6%): post-hoc analysis revealed that this came from 5 control contexts for which the accuracy was particularly low, thus not reliable controls. We decided to remove these 5 controls from our exclusion criteria, as they were particularly difficult, and probably do not indicate that subjects were not doing the task correctly. After restricting to the 15 remaining controls, mean accuracy on controls was 90.0%, showing that participants were not answering randomly.

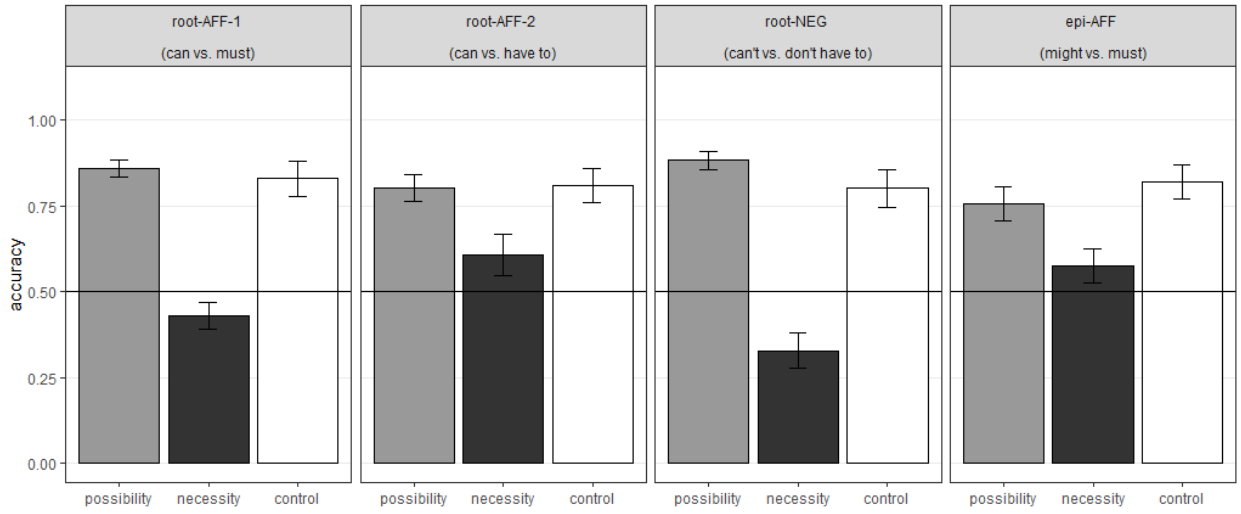
with Subject and Item as random factors, comparing a model with Force as fixed effect to a model without it (Maximal model: Accuracy~Force+(Force|Subject)+(1|Item)). The difference is significant in all conditions, with higher accuracy for possibility modals (Model comparison: all: $\chi^2(1)=20.49$, $p=5.9e\text{-}6$***; ROOT-AFF-1: $\chi^2(1)=60.4$ $p=7.7e\text{-}15$***; ROOT-AFF-2: $\chi^2(1)=7.37$ $p=0.0066$**; ROOT-NEG: $\chi^2(1)=38.1$, $p=6.6e\text{-}10$***; EPI-AFF: $\chi^2(1)=7.93$ $p=0.0048$**). **Negation** — We compared ROOT-AFF-2 and ROOT-NEG, as they included the same lemmas. We used the same method, comparing a model with Negation as fixed effect to a model without it (Maximal model: Accuracy~Negation+(1|Subject)+(1|Item); for the interaction: Accuracy~Force*Negation+(1|Subject)+(1|Item)). We find a significant difference (*can* vs. *can't*: $\chi^2(1)=3.65$, $p=0.056$*; *have to* vs. *not-have-to*: $\chi^2(1)=6.74$, $p=0.0093$**; Interaction Force*Neg: Accuracy~Force*Group+(1|Subject)+(1|Item): $\chi^2(1)=9.24$, $p=0.0024$**). **Flavor** — There was no effect of flavor ($\chi^2(1)=0.14$, $p=0.71$). **Age** (adult vs. child productions) — We then tested for the effect of Age (adult vs child usage), first overall and then for each subcondition (Accuracy~Experiment+(1|Subject)+(1|Item)). We find a significant difference, with an overall lower accuracy for child usage ($\chi^2(1)=260.5$, $p<.001$***). Among possibility conditions, only ROOT-AFF-1 is significant; among necessity conditions, all comparisons are significant except EPI-AFF (**Table 9**). We find a strong interaction Force*Age: the difference in accuracy between possibility and necessity modals for child productions is larger than for adult productions ($\chi^2(1)=46.2$, $p=1.06e\text{-}11$***).

	Mean accuracy[46] (se)		Exact binomial tests (two-sided)	
	possibility	necessity	Possibility	necessity
ROOT-AFF-1	85.1% (0.026)	42.6% (0.039)	p<.001*** 95% CI [0.82, 0.88]	p<.001*** 95% CI [0.39, 0.46]
ROOT-AFF-2	79.6% (0.041)	60.2% (0.060)	p<.001*** 95% CI [0.77, 0.83]	p<.001*** 95% CI [0.56, 0.63]
ROOT-NEG	88.2% (0.027)	32.3% (0.050)	p<.001*** 95% CI [0.86, 0.91]	p<.001*** 95% CI [0.29, 0.36]
EPI-AFF	75.6% (0.050)	56.8% (0.047)	p<.001*** 95% CI [0.73, 0.80]	p<.001*** 95% CI [0.53, 0.61]
Total	82.1% (0.019)	50.1% (0.028)		
ALL	67.4% (0.021)			

Table 8 Accuracy rates and binomial tests by condition, testing for significance versus chance (Experiment 5: children's productions) (n=273, 10 observations per cell)

Fig. 10 Accuracy by condition, Experiment 5, children's productions (n=273)

[46] Accuracy corresponds to the mean accuracy across the 20 contexts initially extracted for each condition. On average, each context was seen by 34.7 participants (ranging between 24 and 47).



	possibility	necessity
ROOT-AFF-1	*$\chi^2(1)=3.12$, p=0.078 (NS)*	*$\chi^2(1)=35.8$, p=2.1e-09****
ROOT-AFF-2	*$\chi^2(1)=5.80$, p=0.016**	*$\chi^2(1)=51.8$, p=6.3e-13****
ROOT-NEG	*$\chi^2(1)=2.78$, p=0.096 (NS)*	*$\chi^2(1)=21.1$, p=4.4e-06****
EPI-AFF	*$\chi^2(1)=3.76$, p=0.053 (NS)*	*$\chi^2(1)=0.22$, p=0.64 (NS)*
all	*$\chi^2(1)=15.9$, p=6.7e-05****	*$\chi^2(1)=175.7$, p<.001****
ALL	*$\chi^2(1)=231.4$, p<2.2e-16****	

Table 9 Results of the model testing effect of Age (adult usage vs. child usage)

3.3 Discussion

Even if participants are less accurate than when judging adults' modals, they are good at identifying possibility modals used by children, for both flavors (mean accuracy on all possibility modals: 82.1%, vs. 87.5% when judging adult modals). Participants' performance for necessity modals is much lower (only 50.1%, vs. 72.3% for adult modals), especially for negated uses (32.3%). Examples like (13) and (14), which led to particularly low accuracy, illustrate children's non adult-like uses of necessity modals with and without negation (and confirm that the effect is likely not just due to participants expecting children to use possibility modals more often than necessity modals).

(13) […] CHILD: what shall I put first?
CHILD: that.
CHILD: what's that?
MOTHER: pardon?

CHILD: I **have to** see a cat. (Becky, 2;08.16) (HSP accuracy: 2.9%)

(14) CHILD: ... no eggs.
MOTHER: I thought we had all of these eggs.
CHILD: they not.
CHILD: they go in the bag.
CHILD: they going in there.
CHILD: they go in there.
MOTHER: oh you're putting them back in there now, are you?
CHILD: you **don't have to** eat them. (Carl, 2;8.07) (HSP accuracy: 20.0%)

Together, these results suggest that children master possibility modals early, as they use them in an adult-like way. Their necessity modals, however, seem delayed: children do not use them in an adult-like way, suggesting that they either haven't mastered their underlying force yet, or that they have difficulty deploying them in the right situations. But if the difficulty with necessity modals we observe here for 2-to-3-year-olds persists into the preschool years, it could explain both types of over-acceptance found in comprehension studies: children would accept necessity modals in possibility contexts because they haven't mastered their underlying force, and accept possibility modals in contexts where adults prefer necessity modals because they lack a stronger alternative (i.e., they have not yet worked out scale-mate relations for English modals).

Importantly, we find no evidence in favor of a necessity bias. Children's highly adult-like uses of possibility modals might even suggest a bias towards possibility. Note that this lack of evidence doesn't necessarily entail that children don't rely on a necessity bias when acquiring modals.[47] It's conceivable that children use the bias to acquire necessity modals, but fail to use them in an adult-like way for independent reasons, as alluded to above. However, the lack of straightforward evidence for a necessity bias in our child results, together with its superfluity given our input results, suggests that a bias towards strong meanings is dispensable, even for modals.

4 General discussion and future directions

How do children figure out the force of their modals? In particular, what prevents them from falling prey to the subset problem modals give rise to, and hypothesizing possibility meanings for necessity modals? To address these questions, we examined the modals that English-learning young children get exposed to and produce themselves. We find that children seem to master possibility modals early: already at age 2, they use them productively, with and without negation, and in an apparently adult-like ways. Children, however, seem to struggle with necessity modals. The few necessity modals they produce do not seem adult-like, and appear in situations where

[47] As pointed out by an anonymous reviewer, our results may also be consistent with theories of acquisition that use a necessity bias that is sensitive to input frequencies, along the lines of Piantadosi et al.'s (2012b) Bayesian learner (see also Piantadosi et al. 2012a). We intend to address this in future work.

adults would prefer possibility modals. If this struggle with necessity modals persists into the preschool years, it could explain why prior studies show that children tend to accept them in possibility contexts (they're uncertain about their force), and also why they accept possibility modals in necessity contexts (they lack a stronger alternative).

Yet children eventually figure out necessity modals, and the question is how. From our input study, we see that given the way modals are used in speech to children, children cannot reliably make use of downward entailing environments like negation, as Gualmini and Schwarz (2009) proposed as a general solution for subset problems. Negation may even be partly responsible for children's difficulties with necessity modals. First, its scopal behavior with modals is not uniform: some, but not all, necessity modals outscope negation. If children were to rely on negation to figure out the force of necessity modals, they could be misled into thinking that a modal like *must* expresses possibility (*must not ~ cannot*), if they assume that negation scopes over the modal. Second, we find that negated necessity modals are rare in speech to children, perhaps for functional reasons, as speakers can express non-necessity via scalar implicatures triggered by the simple use of a possibility modal. Finally, Experiment 1 shows that the context is the least clear about force for negated necessity modals. This seems to be due to their use in impossibility situations, as ways to soften requests or orders (e.g., 'you *don't have to* break those things', used as a prohibition). Negation, however, might be quite useful for children to home in early on possibility meanings for possibility modals: It occurs frequently with possibility modals, and our experimental results suggest that negated possibility contexts are particularly informative about the force of the modal.

If learners can't rely on downward entailing environments to solve the subset problem for modals, might they then need a necessity bias? Such a bias is in principle not necessary, as our experiments show that the conversational context in which modals are used is informative about force. What exactly about the context gives away modal force? Our initial foray into contextual features suggests one factor that could be particularly helpful for deontic modals is the perceived (un)desirability of the prejacent (e.g., 'you have to eat your peas' vs. 'you can have a cookie'). Desirability is a notion that should uncontroversially be available to young children, though its association with obligations may take some time. Experiment 4 confirms the potential usefulness of desirability, and shows that root *necessity* modals tend to occur with undesirable prejacents, and *possibility* modals with desirable prejacents. Participants are better at guessing necessity modals when they occur with undesirable prejacents. For epistemic modality, our post-hoc analysis suggests that contexts that explicitly highlight salient evidence in favor of the prejacent may bias interpretations towards necessity. Other aspects of the context could also prove useful, including situational cues (e.g., who the interlocutors are), cues from world knowledge (e.g., what is allowed or prohibited), or pragmatic cues (what the speaker is trying to achieve, in particular performing orders, permissions or prohibitions). We plan to explore this further in future work.

Taken together, however, the results from our input study and our child study seem to lead to a conundrum: if the conversational context is informative for both forces (Experiment 2 and 3 suggest that context is particularly helpful for necessity modals), why should children struggle

more with necessity modals? Several factors could be at play. First, these difficulties might be a matter of quantity, rather than quality, as our corpus results show that necessity modals are less common in the input.[48] Second, children may need to accrue some world knowledge to be able to make use of contextual cues (e.g., the association between obligations and undesirability). Children might also face conflicting cues. In particular, negative contexts seem useful for possibility modals, but misleading for some necessity modals. As discussed, if children assume the same scope relations for *can* and *must*, uses of *mustn't* could suggest to them that it expresses possibility. And interestingly, when we compare our two affirmative root conditions, we find a lower performance for children's *must* than for *have to* (*must*: 42.6% vs. *have to*: 60.2%), though this could reflect differences in their input frequency (*have to* represents 12% of all parents' modal utterances, *must* only 2.3%). The interaction with negation could also explain why possibility modals are mastered so early: our study of the input shows that negated possibility modals are frequent, and used in particularly informative contexts (e.g. to talk about prohibitions or physical impossibilities). Last, children's struggles with necessity modals might not reflect a lack of knowledge of their underlying force, but difficulty deploying this knowledge in the right situations. Children may have trouble tracking what information is shared amongst interlocutors, and thus grasping the intended domain of quantification (as has been argued for definite descriptions, where children seem to be overly permissive in using and accepting definites in contexts in which their uniqueness presupposition is not satisfied; Karmiloff-Smith 1979; Schaeffer and Matthewson 2005; Brockmann et al. 2018).

What we hope to have shown here is that the conversational context is a rich source of information for modal force, so that, in principle, learners can solve the subset problem without having to rely on potentially misleading negative contexts, nor on a controversial strong (necessity) bias. We have only begun to scratch the surface for which properties of the context are useful and why, but our results show that even the short, written snippets of conversation we provided to adult participants are informative, and this informativity is above and beyond the more limited context in the prejacent alone (Exp 1 vs. Exp 2/3). In fact, even just the information in the prejacent leads to above chance success in Experiment 2. Note that the context in real life is far richer than in Experiment 1 contexts, which lack prosodic and emotive information carried by speech, shared knowledge between mother and children, and visual information (among others). Our studies, in combination with the fact that real life contexts are much richer, suggest that usage contexts may richly support modal force learning. That said, we don't know yet how representative these contextually useful cases are in children's actual experience, and if the additional factors (e.g., prosody) will be enriching, as seems intuitive, or obfuscating. Now that we are confident that information of some kind is available in the context, and that desirability and mention of evidence are two promising types of information supporting modal meanings, we can pursue research asking

[48] We do not have enough data from each child to see whether children's adult-like uses of necessity modals correlate with frequency in their input (i.e., whether children who hear more necessity modals use them more appropriately) but intend to pursue this question in future work. We thank an anonymous reviewer for this great suggestion. See Dieuleveut 2021a for preliminary results and discussion.

whether children actually make use of these (and other) specific aspects of the context to acquire modal force.

Our child study shows no evidence for a necessity bias in children's early modal productions. In fact, children's early successes with possibility modals and failures with necessity modals could even suggest a bias towards possibility. Still, our results cannot rule out that children actually have a necessity bias, and grasp the force of necessity modals, but fail to use them in an adult-like way for independent reasons. In future work, we plan to test for such biases more directly through a novel modal word task, adapted from an adult study in Dieuleveut *et al.* (2021b), to see what meanings children attribute to novel modals.

To address what aspects of the context are useful, and whether children actually make use of them in their modal force acquisition, we plan on testing whether various features of the context in the input are good predictors of children's mastery of necessity modals, as indexed by accuracy on the child HSP task. For instance, to see whether desirability actually matters in children's (root) modal acquisition, we could test whether frequent uses of necessity modals with undesirable prejacents in the input predict earlier mastery of necessity modals: will a child whose parents primarily use necessity modals with undesirable prejacents use necessity modals in an adult-like way sooner than a child whose parents use necessity modals more often with desirable prejacents? (see Dieuleveut 2021 for preliminary results).[49]

Before we conclude, we would like to briefly discuss potential implications of our findings for how children acquire modal force in languages beyond English, and in particular, in languages with 'variable force' modals. As discussed in section 1.1, in a language like English where modals come in both forces, we can expect speakers to use possibility and necessity modals in fairly distinct situations, and notably, to avoid using possibility modals in necessity situations (modulo politeness considerations). And indeed, our input results show that speakers use possibility and necessity modals in distinct situations that are highly reflective of force. But in a language that lacks modal duals, speakers are more likely to use particular modals in both possibility and necessity situations. For variable force modals that are underlying possibility modals, like Nez Perce *o'qa*, it seems like negation would thus be crucial for learners to home in on its underlying force—just like it was for Deal (2011) to argue for a possibility analysis. For variable force modals that are underlying necessity modals as in St'át'imcets or Washo, the challenge may be much greater. Not only might speakers use the same modals in possibility and necessity situations, but learners may not be able to rely on negation, given that it can't scope over modals in these languages. And yet speakers seem to have converged on necessity meanings for these modals, as evidenced by their preferred translations using English necessity modals (Rullman and Matthewson 2018, Bochnak 2015). Here fieldworkers can and do rely on such translations as evidence for the modals' underlying force, but this strategy is obviously inaccessible to the child. How do learners figure out their underlying force? This situation might at first blush argue for a

[49] For a similar investigation of what aspects of the input predict children's understanding of *think* and *know*, see Dudley (2017).

necessity bias. However, it could also be that while these modals can in principle be used in possibility situations, in practice, variable force modals are mostly used in contexts where English speakers use necessity modals, in which case, their acquisition could involve the same reliance on contextual cues that we've proposed for the acquisition of English modals.

Conclusion

This study has laid the groundwork for understanding when and how children figure out the force of the modals in their language, and in particular, how they may solve the subset problem of modals, by looking at young children's natural productions and their input. Our child results suggest that at age two children have mastered possibility modals, but they struggle with necessity modals. Our input results suggest that, in principle, learners could learn force simply based on how modals are used, as the conversational context is informative about force. If children are able to pick up on these conversational cues, they could figure out modal force without having to rely on negation—which we've shown is particularly unreliable for necessity modals, nor on a bias towards necessity—a bias for which we find no evidence in child productions. Our results are thus in line with recent discussions of other subset problems arguing that strong meaning biases may have no place in acquisition (Musolino et al. 2019).

References

Abbott, B. (2003). A reply to Szabó's "Descriptions and Uniqueness". *Philosophical Studies*, *113*(3), 223-231.

Acredolo, C., & Horobin, K. (1987). Development of relational reasoning and avoidance of premature closure. *Developmental Psychology*, *23*(1), 13.

Austin, J. L. (1975). *How to do things with words*. Oxford university press.

Barner, D., & Bachrach, A. (2010). Inference and exact numerical representation in early language development. *Cognitive psychology*, *60*(1), 40-62.

Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: The role of scalar alternatives in children's pragmatic inference. *Cognition*, *118*(1), 84-93.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255-278.

Berwick, R. C. (1985). *The acquisition of syntactic knowledge* (Vol. 16). MIT press.

Bindra, D., Clarke, K. A., & Shultz, T. R. (1980). Understanding predictive relations of necessity and sufficiency in formally equivalent" causal" and" logical" problems. *JEP: General*, *109*(4), 422.

Bochnak, M. R. (2015). Underspecified modality in Washo. In *Proceedings of the eighteenth and nineteenth workshop on structure and constituency in the languages of the Americas* (pp. 3-17).

Bochnak, M. R. (2015). Variable force modality in Washo. In *Proceedings of North-East Linguistics Society* (Vol. 45, pp. 105-114).

Brockmann, S., McConnell, S., Hacquard, V., & Lidz, J. (2018). Children's comprehension of pronouns and definites. *ZAS Papers in Linguistics*, *60*, 259–276.

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage* (Vol. 4). Cambridge university press.

Byrnes, J. P., & Duff, M. A. (1989). Young children's comprehension of modal expressions. *Cognitive Development*, *4*(4), 369-387.

Cariani, F., & Santorio, P. (2017). Will done better: Selection semantics, future credence, and indeterminacy. *Mind*, *127*(505), 129-165.

Chierchia, G., Crain, S., Guasti, M. T., Gualmini, A., & Meroni, L. (2001). The acquisition of disjunction: Evidence for a grammatical view of scalar implicatures. In *Proceedings of the 25th Boston University conference on language development*(pp. 157-168).

Cournane, A. (2015). Revisiting the epistemic gap: Evidence for a grammatical source. In *Proceedings of the 39th annual Boston University conference on language development*.

Cournane, A. (2015). *Modal development: input-divergent L1 acquisition in the direction of diachronic reanalysis* (Doctoral dissertation).

Crain, S., Ni, W., & Conway, L. (1994). Learning, parsing and modularity. *Perspectives on sentence processing*, 443-467.

De Haan, F. (2013). *The interaction of modality and negation: A typological study*. Routledge.

Deal, A. R. (2011). Modals without scales. *Language*, *87*(3), 559-585.

Dieuleveut, A. (2021). *Finding modal force* (Doctoral dissertation).

Dieuleveut, A., van Dooren, A., Cournane, A., & Hacquard, V. (2019). Learning modal force: evidence from children's production and input. In *Proceedings of the 2019 Amsterdam Colloquium* (pp. 111-122).

Dieuleveut, A., van Dooren, A., Cournane, A., & Hacquard, V. (2019). Acquiring the Force of Modals: Sig You Guess What Sig Means?. In *Proceedings of the 42nd Annual to the Boston University Conference on language development (BUCLD42)*.

Drubig, H. B. (2001). On the syntactic form of epistemic modality. *ms., University of Tübingen*.

Dudley, R. E. (2017). *The role of input in discovering presuppositions triggers: Figuring out what everybody already knew* (Doctoral dissertation).

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, *73*(2), 135-176.

Gleitman, L. R., & Gillette, J. (2017). The role of syntax in verb learning. *The handbook of child language*, 413-427.

Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. C. (2005). Hard words. *Language learning and development*, *1*(1), 23-64.

Gleitman, L. R., Gleitman, H., Landau, B., & Wanner, E. (1989). Where learning begins: Initial representations for language learning. *Language: Psychological and biological aspects*, 150-193.

Grice, H. P., Cole, P., & Morgan, J. L. (1975). Logic and conversation. *1975*, 41-58.

Gualmini, A., & Schwarz, B. (2009). Solving learnability problems in the acquisition of semantics. *Journal of Semantics*, *26*(2), 185-215.

Hackl, M. (1998). On the Semantics of" Ability Attributions".

Hacquard, V. (2011) Modality. In C. Maienborn, K. von Heusinger, P. Portner (eds.) *Semantics: An International Handbook of Natural Language Meaning. HSK 33.2* Berlin: Mouton de Gruyter. 1484-1515.

Halberda, J. (2006). Is this a dax which I see before me? Use of the logical argument disjunctive syllogism supports word-learning in children and adults. *Cognitive psychology*, *53*(4), 310-344.

Hirst, W., & Weil, J. (1982). Acquisition of epistemic and deontic meaning of modals. *Journal of child language*, *9*(3), 659-666.

Hoffmann, T. (1966). Past tense replacement and the English modal auxiliary system. *Harvard computation laboratory report NSF*, *17*.

Horn, L. R. (1972). On the semantic properties of logical operators. Doctoral dissertation*, UCLA*.

Iatridou, S., & Zeijlstra, H. (2013). Negation, polarity, and deontic modals. *Linguistic inquiry*, *44*(4), 529-568.

Jeretič, P. (2018). Evidence for children's dispreference for weakness: A corpus study. Ms. New York, NY: New York University.

Karmiloff-Smith, A. (1979). Micro-and macrodevelopmental changes in language acquisition and other representational systems. Cognitive science, 3(2), 91-117.

Klecha, P. (2016). Modality and embedded temporal operators. *Semantics and Pragmatics*, *9*, 9-1.

Koring, L., Meroni, L., & Moscati, V. (2018). Strong and Weak Readings in the Domain of Worlds: A Negative Polar Modal and Children's Scope Assignment. *Journal of psycholinguistic research*, *47*(6), 1193-1217.

Kratzer, A. (1981). Partition and revision: The semantics of counterfactuals. *Journal of Philosophical Logic*, *10*(2), 201-216.

Kratzer, A. (1991). Modality. *Semantics: An international handbook of contemporary research*, *7*, 639-650.

Kratzer, A. (2012). *Modals and conditionals: New and revised perspectives* (Vol. 36). Oxford University Press.

Kuczaj, S. A., & Maratsos, M. P. (1975). What children can say before they will. *Merrill-Palmer Quarterly*.

Landau, B., & Gleitman, L. R. (1985). Cognitive science series, 8. *Language and experience: Evidence from the blind child. Cambridge, MA, US: Harvard University Press*.

Lin, J., Weerman, F., & Zeijlstra, H. (2015). Emerging NPIs: The acquisition of Dutch hoeven 'need'. *The Linguistic Review*, *32*(2), 333-374.

Lyons, J. (1977). *Semantics* (Vol. 2). Cambridge University Press.

MacWhinney, B. (2000). *The CHILDES project: The database* (Vol. 2). Psychology Press.

Mandelkern, M., Schultheis, G., & Boylan, D. (2015). I believe I can φ.

Moscati, V., & Crain, S. (2014). When negation and epistemic modality combine: the role of information strength in child language. *Language Learning and Development*, *10*(4), 345-380.

Moscati, V., Zhan, L., & Zhou, P. (2017). Children's on-line processing of epistemic modals. *Journal of child language*, *44*(5), 1025-1040.

Musolino, J. (1999). Universal Grammar and the acquisition of semantic knowledge: An experimental investigation into the acquisition of quantifier-negation interaction in English.

Musolino, J., Crain, S., & Thornton, R. (2000). Navigating negative quantificational space. *Linguistics*, *38*(1), 1-32.

Musolino, J., Laity d'Agostino, K., & Piantadosi, S. (2019). Why We Should Abandon the Semantic Subset Principle. *Language Learning and Development*, *15*(1), 32-46.

Noveck, I. A. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, *78*(2), 165–188.

Ozturk, O., & Papafragou, A. (2015). The acquisition of epistemic modality: From semantic meaning to pragmatic interpretation. *Language Learning and Development*, *11*(3), 191-214.

Papafragou, A. (1998). Inference and word meaning: The case of modal auxiliaries. *Lingua*, *105*(1-2), 1-47.

Papafragou, A. (1998). The acquisition of modality: Implications for theories of semantic representation. *Mind & language*, *13*(3), 370-399.

Peterson, T. R. G. (2010). *Epistemic modality and evidentiality in Gitksan at the semantics-pragmatics interface* (Doctoral dissertation, University of British Columbia).

Piantadosi, S. T. (2011). *Learning and the language of thought* (Doctoral dissertation, Massachusetts Institute of Technology).

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2013). Modeling the acquisition of quantifier semantics: a case study in function word learnability. *Under review*.

Piéraut-Le Bonniec, G. (1980). The Development of Modal Reasoning Genesis of Necessity and Possibility Notions.

Portner, P. (2009). *Modality* (Vol. 1). Oxford University Press.

Rakoczy, H., Warneken, F., & Tomasello, M. (2007). "This way!","No! That way!"—3-year olds know that two people can have mutually incompatible desires. *Cognitive Development*, *22*(1), 47-68.

Rasin, E., & Aravind, A. (2020). The nature of the semantic stimulus: the acquisition of every as a case study. *Natural Language Semantics*, 1-37.

Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: evidence from 14-and 18-month-olds. *Developmental psychology*, *33*(1), 12.

Robinson, E. J., Rowley, M. G., Beck, S. R., Carroll, D. J., & Apperly, I. A. (2006). Children's sensitivity to their own relative ignorance: Handling of possibilities under epistemic and physical uncertainty. *Child development*, *77*(6), 1642-1655.

Ruffman, T., Aitken, J., Wilson, A., Puri, A., & Taumoepeau, M. (2018). A re-examination of the broccoli task: Implications for children's understanding of subjective desire. *Cognitive Development*, *46*, 79-85.

Rullmann, H., Matthewson, L., & Davis, H. (2008). Modals as distributive indefinites. *Natural Language Semantics*, *16*(4), 317-357.

Schaeffer, J., & Matthewson, L. (2005). Grammar and pragmatics in the acquisition of article systems. Natural Language & Linguistic Theory, 23(1), 53-101.

Searle, J. R., & Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language* (Vol. 626). Cambridge university press.

Skordos, D., & Papafragou, A. (2016). Children's derivation of scalar implicatures: Alternatives and relevance. *Cognition*, *153*, 6-18.

Snedeker, J. (2000). Cross-situational observation and the semantic bootstrapping hypothesis. In *Proceedings of the thirtieth annual child language research forum. Stanford, CA: Center for the Study of Language and Information*.

Snedeker, J., Gleitman, L., & Brent, M. (1999). The successes and failures of word-to-world mapping. In *In*.

Stalnaker, R. (1978). Assertion.

Stromswold, K. J. (1990). *Learnability and the acquisition of auxiliaries*, Doctoral dissertation, MIT).

Team, R. C. (2013). R: A language and environment for statistical computing.

Theakston, A. L., Lieven, E. V., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of child language*, *28*(1), 127-152.

Van der Auwera, J., & Plungian, V. A. (1998). Modality's semantic map.

van Dooren, A., Dieuleveut, A., Cournane, A., & Hacquard, V. (2017). Learning what must and can must and can mean. In *Proceedings of the 21st Amsterdam Colloquium*.

Von Fintel, K., & Iatridou, S. (2008). How to say ought in foreign: The composition of weak necessity modals. In *Time and modality* (pp. 115-141). Springer, Dordrecht.

Wellman, H. M., & Woolley, J. D. (1990). From simple desires to ordinary beliefs: The early development of everyday psychology. *Cognition*, *35*(3), 245-275.

White, A. S., Hacquard, V., & Lidz, J. (2018). Semantic information and the syntax of propositional attitude verbs. *Cognitive science*, *42*(2), 416-456.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological review*, *114*(2), 245.

Yanovich, I. (2016). Old English* motan, variable-force modality, and the presupposition of inevitable actualization. *Language*, *92*(3), 489-521.

Zhou, P., & Crain, S. (2009). Scope assignment in child language: Evidence from the acquisition of Chinese. *Lingua*, *119*(7), 973-988.