Communicative pressures shape language during communication (not learning): Evidence from casemarking in artificial languages

Kenny Smith[a], Jennifer Culbertson[a]

[a]*Centre for Language Evolution, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, 3 Charles Street, Edinburgh, EH8 9AD, United Kingdom*

Abstract

Natural languages are designed for efficient communication. A classic example is Differential Case Marking, when nouns are marked for their grammatical role only if this information cannot be derived from world knowledge (e.g. only atypical objects need to be linguistically marked as objects). Fedzechkina et al. (2012) present experimental evidence from an artificial language learning paradigm suggesting that biases in learning favour Differential Case Marking: learners exposed to a language with optional casemarking restructure the input, using casemarkers more in situations where marking would reduce ambiguity, despite the fact that they never use the artificial language in a communicative task where ambiguity matters. This is surprising given previous studies suggesting that biases in learning favour simplicity and are agnostic with respect to communicative function. We present 3 experiments investigating whether biases for communicatively-efficient Differential Case Marking exist in learning. Contrary to Fedzechkina et al. (2012), we find no evidence for such a bias in learning: participants' use of case is impervious to the ambiguity of unmarked objects, and modulated by statistical properties of their input, observations which are inconsistent with their hypothesis and better explained by biases early in learning favouring iconicity in the form-meaning mapping, e.g. whereby atypical meanings are associated with atypical forms. However, we find good evidence that participants' behaviour in actual communicative interaction *are* driven by efficient communication considerations: in interaction participants exhibited the expected Differential Object Marking pattern. This suggests that languages adapt to communicative efficiency constraints as a result of being used in communication, rather than due to biases in human learning favouring communicatively-efficient languages.

Key words: language universals; language evolution; learning biases; efficient communication; iconicity; artificial language learning

1. Introduction

Natural languages seem designed to be communicatively *efficient*: they appear to optimally trade off communicative function, which would push us to be maximally explicit, and effort, which would push us to say a little as we can get away with (e.g., Zipf, 1936, 1949; Comrie, 1989; Jäger, 2007). For instance, the lexicons of natural languages seem to be

Email address: kenny.smith@ed.ac.uk (Kenny Smith)

designed so that words we use frequently are short, whereas lower-frequency words tend to be longer (Zipf, 1936). This is more efficient than if frequent words were long and rare words were short. Furthermore, short words tend to have multiple senses (i.e. they are potentially ambiguous, Piantadosi et al., 2012); since that ambiguity is resolved in context, re-using low-effort short words in this way is efficient.

In this paper we focus on a particular instance of apparent communicative efficiency known as *Differential Case Marking*. All languages provide a means for indicating the role of participants in events, i.e. who did what to who. While this is often accomplished in whole or in part by word order, some languages additionally or alternatively use adpositions or affixes to perform this function, for example marking the subject or object of a sentence with an additional affix indicating its role (see examples (1)–(5) below). Casemarking involves effort (the production and processing of the marker), and yet some of the information encoded by the case marker may be recoverable from context: if I tell you that there was a painting event involving a person and a wall, you are likely to infer that the person painted the wall rather than the reverse, even if I provide no explicit information indicating who did what to who, because humans are highly typical *agents* of actions and walls or other inanimate objects are highly typical *undergoers* of actions carried out by others (i.e., they are highly atypical agents). Languages with Differential Case Marking systems exploit this fact: typical event participants can go unmarked, whereas atypical event participants are more likely to have their role in the event explicitly casemarked (e.g. Silverstein, 1976; Bossong, 1991; Dixon, 1979; see Witzlack-Makarevich & Seržant, 2018 for review); in the example above, in a Differential Case Marking language casemarking would only be required if a wall somehow painted a person, in which case at least one of the nouns would be marked to explicitly indicate its atypical or surprising role. This produces a communicatively-efficient configuration, and one widely-accepted explanation (e.g. Comrie, 1989) is therefore that Differential Case Marking is the product of a trade-off between ambiguity avoidance and economy of effort, since marking is restricted to events that are potentially ambiguous.

This approach to explaining Differential Case Marking therefore resides in language use, maximising communication (i.e. minimising ambiguity) while minimising effort (here, case marking). However, at least one experimental study (Fedzechkina et al., 2012, henceforth FNJ) provides evidence that a bias for efficient casemarking exists in learning: adult participants trained on an artificial language with optional casemarking (such that case is sometimes marked and sometimes not, randomly and without being conditioned on the typicality of event participants) produce a language in which casemarking is conditioned on typicality. Crucially, this occurs despite the fact that those participants never use those languages for communication. While there are several other papers reporting biases in learning which favour communicatively-optimal configurations (Carstensen et al., 2015; Fedzechkina et al., 2017, 2018; Levshina, 2018; Roberts & Fedzechkina, 2018; Kurumada & Grimm, 2019; Fedzechkina & Jaeger, 2020), another body of other work using similar methods suggests that biases operating in learning are at best agnostic with respect to communicative function and often actively erode communicative utility in favour of other factors such as increasing representational simplicity (e.g. Kirby et al., 2008, 2015; Silvey et al., 2015; Carr et al., 2017, 2020; Smith et al., 2020). From this perspective, FNJ's result that biases in learning favour communicative efficiency is surprising.

Here we present a series of artificial language experiments exploring whether biases in

learning indeed favour communicative efficiency, building on FNJ's work on Differential Case Marking. To preview briefly, we fail to replicate FNJ's results, and across 3 experiments find little evidence of a bias favouring Differential Case Marking in learning. The (rather weak) evidence we see for biases in learning is instead consistent with a bias for iconicity, where atypical arguments receive atypical marking regardless of whether this would facilitate or hamper communication. However, when our participants use the artificial language in a communicative task we see a rapid and reliable shift towards communicatively-efficient Differential Case Marking: participants *can* spontaneously create Differential Case Marking systems, but do so only when using the language to communicate, where ambiguity avoidance becomes relevant. This casts doubt on FNJ's account that Differential Case Marking is a product of biases in learning favouring communicative efficiency. Instead, it supports what we regard as the more intuitive view that such biases operate during actual communication. In other words, languages are adapted for communicative efficiency during communication, not learning.

1.1. Differential Case Marking

Before turning to our experiments, we first describe Differential Case Marking in more detail and outline several alternative accounts which have been proposed to explain this phenomenon.

As outlined above, many language employ grammatical devices other than word order for *argument marking*. These including *flagging* on the arguments (event participants) themselves by means of adpositions or case markers (see examples (1)–(2) from Hungarian and German, where case is marked on the noun or the article respectively, leaving word order free to vary and convey other information, e.g. emphasis).[1]

(1) Hungarian (Finno-Ugric)

a. a kerékpáros az autó-t elkerülte
the cyclist the car-ACC avoided
'The cyclist avoided the car'

b. az autó-t a kerékpáros elkerülte
the car-ACC the cyclist avoided
'The cyclist avoided the car'

c. a kerékpáros-t az autó elkerülte
the cyclist-ACC the car avoided
'The car avoided the cyclist'

d. az autó a kerékpáros-t elkerülte
the car the cyclist-ACC avoided

[1]Case is glossed using capitals, e.g. ACC is the gloss for an accusative marker or form. Asterisks indicate ungrammaticality and brackets indicate optionality, e.g. (-affix) indicates that affix is optional, (*-affix) indicates that the sentence becomes ungrammatical if the affix is omitted.

'The car avoided the cyclist'

(2) German (Indo-European)

a. Der Radfahrer mied das Auto
the.NOM cyclist avoided the.ACC car
'The cyclist avoided the car'

b. Das Auto mied der Radfahrer
the.ACC car avoided the.NOM cyclist
'The cyclist avoided the car'

c. Das Auto mied den Radfahrer
the.NOM car avoided the.ACC cyclist
'The car avoided the cyclist'

d. Den Radfahrer mied das Auto
the.ACC cyclist avoided the.NOM car
'The car avoided the cyclist'

Our focus in this paper is on instances where marking appears to be deployed in a communicatively-efficient manner. In a number of languages, marking does not apply to all arguments, but is contingent on properties of the argument nouns (Silverstein, 1976; Bossong, 1991; Dixon, 1979), with some nouns being left unmarked. This differential argument marking is often conditioned on animacy, with atypical arguments being more likely to be marked.[2] Typical subjects[3] are animate and indeed human; thus in Differential Subject Marking systems, case-marking may be dropped on animate/human subjects, whereas atypical subjects are more likely to be explicitly marked, as in example (3) (where a pig is a less typical subject than a man). Typical objects are inanimate; thus in Differential Object Marking systems, we see the inverse pattern, with *animate* objects more likely to be marked (examples (4)–(5)): in Camling only human (high animacy) objects can be marked with the dative case; similarly, in Spanish the preposition *a* is used to mark human objects.[4]

[2]In many languages, casemarking is also dependent on the definiteness or specificity of the argument in question. We do not address this further here, however note that as for animacy, arguments with atypical definiteness or specificity values relative to their argument type are more likely to be marked. For example, typical subjects are not just animate, but also definite and specific — they are unique and identifiable in the context — while typical objects are inanimate, indefinite, and non-specific (see Jäger, 2007, for related corpus evidence).

[3]We will use the terms "subject" and "object" throughout, which refer to the syntactic roles of event participants; note that in the literature on Differential Case Marking the terms "agent" and "patient" are often used instead to refer to the semantic roles of participants.

[4]It should be noted that despite being described as "extremely widespread" (Aissen, 2003, p. 439) or "a universal tendency" (Haspelmath, 2018), several recent papers have questioned the strength of evidence for Differential Case Marking as a truly robust cross-linguistic universal. For example, Bickel & Witzlack-Makarevich (2008) look across 333 languages for predictive relationships between the probability of case-

(3) Fore (Papuan, example from de Hoop & Malchukov, 2008, p569)

a. Yagaa wá aegúye
pig man hit
'The man hits the pig'

b. Yagaa-wama wá aeg'uye
pig-ERG man hit
'The pig hits the man'

(4) Camling (Sino-Tibetan, example from Kittilä, 2005, p506)

a. khu-wa lungto-wa pucho(*-lai) set-yu
he-ERG stone-INSTR snake(*-DAT) kill-3
'He killed the snake with a stone'

b. khana khut(-lai) ta-set-yu
I he-(DAT) 2-kill-3
'You killed him'

(5) Spanish (Indo-European)

a. Dani besó a la mujer
Dani kissed to the woman
'Dani kissed the woman'

b. Dani besó la imagen
Dani kissed the picture
'Dani kissed the picture'

---

marking and animacy and definiteness scales known to be relevant for Differential Case Marking. They find evidence of such relationships, but only for two or three independent groups of languages. Along similar lines, Sinnemäki (2014) surveys 744 languages for evidence of Differential Object Marking and finds that roughly as many independent groupings of languages have object-marking conditioned on animacy/definiteness as do not. Bickel & Witzlack-Makarevich (2008) and Sinnemäki (2014) both conclude that differential marking may reflect features of historical development peculiar to some groups of related languages, rather than reflecting universal biases in learning and/or communication. However, Sinnemäki (2014) finds that among systems of object marking that condition marking on *some* semantic feature(s) of arguments, those based on animacy (and definiteness) are by far the most common (with less-communicatively relevant features like grammatical gender and number being less common). In addition, Börstell (2019) argues that a number of historically independent sign languages (not included in the samples cited above) show Differential Case Marking-like phenomena. To the extent that such phenomena are found in multiple independent sign languages, this may substantially broaden the cross-linguistic evidence for Differential Case Marking, which would in turn suggest that it reflects at least the interplay of historical constraints with a (perhaps relatively weak) universal bias in learning or use (as argued by Seržant, 2018). While the universality of Differential Case Marking systems may have been over-stated, it has thus arisen independently at least twice (and possibly many more times in independent sign languages), and is probably over-represented among restricted casemarking systems. Therefore it merits consideration as a recurring structural configuration for which functional (as well as historical) explanations should be considered.

1.2. Explanations for differential case marking: efficient communication versus iconicity

What leads to these patterns of differential marking? One possibility, advanced by e.g. Comrie (1989) and sketched briefly above, is that Differential Case Marking represents a trade-off between communicative function and efficiency. Using explicit argument marking reduces the possibility of miscommunication, specifically reducing the likelihood of the listener confusing the roles of the arguments in the event being described. This is especially true in languages in which word order is not a reliable cue to participant roles. However, world knowledge will often disambiguate; since inanimates are more likely to be objects than subjects, marking inanimate objects is somewhat redundant, and the marking could be dropped in the interests of minimising effort (either the speaker's effort in producing the marker, or the hearer's effort in processing it); by the same logic, since inanimates are *atypical* subjects, marking them in that role is more likely to be communicatively helpful (and less likely to be redundant), while marking is less necessary for animate subjects. Differential Case Marking therefore represents a potentially optimal trade-off between ambiguity avoidance and economy (Jäger, 2007), marking only when ambiguity poses a greater risk of communication breakdown (e.g. where the object is atypical and risks being misinterpreted as the subject).

A related explanation is that differential marking is not motivated by ambiguity avoidance per se, but represents an example of a more general iconicity preference, a "grand isomorphism" (Givón, 1991), where unusual events/concepts/structures/constituents tend to be associated with special (standardly, more weighty) linguistic material, which Haspelmath (2008) dubs *iconicity of markedness matching*, where the term *marked* does double duty to refer both to atypicality at the conceptual level and weightiness in the surface signal (see also Haspelmath, forthcoming, for the closely related *form-frequency correspondence* proposal). For casemarking, this has be formalised (Aissen, 2003) as the alignment of a scale of grammatical function (where subjects outrank objects) with a scale of animacy (where animates outrank inanimates), in interaction with a preference for overall economy (i.e. penalising casemarking everywhere). This predicts that marking will be dispreferred except for unusual combinations of grammatical function and animacy, which is the pattern seen in Differential Case Marking. The prediction made by iconicity accounts is essentially the same as the ambiguity avoidance account—mark atypical arguments—but the explanation is with reference to iconicity rather than efficient communication.

Given that these two accounts point to very different mechanisms for explaining Differential Case Marking, there has been considerable debate in the linguistics literature as to which is preferable. However, this debate has been on purely theoretical grounds, since data from existing languages appear to be consistent with both, and these in-principle arguments seem relatively weak, making experimental evidence desirable (see also Seržant, 2018, for discussion). For instance, both Aissen (2003) and Haspelmath (forthcoming) point out that differential marking in many languages applies based on the atypicality of the arguments, without reference to their in-the-moment ambiguity or the relative atypicality of a pair of arguments in a given sentence; any argument past a certain critical point on the animacy scale is marked even if ambiguity seems unlikely to be a problem in all such cases, as in example (6) where the preposition *a* is obligatory despite not plausibly being required to disambiguate participant roles.

(6) Spanish (Indo-European)

El asesino asesinó a su víctima
The murderer murdered to their victim

'The murderer murdered his victim'

They take this as evidence that differential marking reflects a bias for iconicity, rather than ambiguity avoidance. However, as pointed out by Seržant (2018), other pressures operating on linguistic systems during their learning and use might lead to a preference for predictable patterns of marking to emerge from initially context-dependent and flexible differential marking. For example, individuals exposed to variation in marking might seek predictable local (i.e. linguistic) factors which govern the appearance or absence of markers. Configurations that are *more likely* to be ambiguous may therefore come to be obligatorily marked (cf. the more general process of obligatorification whereby initially probabilistic, pragmatically-conditioned patterns of variation lose flexibility and become obligatory, Lehmann, 1985; Fehér et al., 2019, or the tendency for initially unconditioned variation to become conditioned on local linguistic context, Smith & Wonnacott, 2010; Smith et al., 2017). Furthermore, it is also worth noting that there are some languages which in fact appear to condition marking on the relative (rather than absolute) animacy of arguments (e.g. de Hoop & Malchukov, 2008 cite Awtuw, a Papuan language, and Fore as examples of this kind) or where speakers can mark or not mark arguments entirely flexibly as appropriate, (Comrie, 1989, p. 130 cites Hua, a Papuan language, as an example of this sort); the flexibility of differential marking might therefore be expected to evolve over time and differ between languages. Adjudicating between communicative efficiency and iconicity accounts of differential marking on purely theoretical grounds therefore seems difficult, making experimental evidence which can arbitrate between mechanisms particularly valuable.

1.3. Fedzechkina et al (2012)

There is a growing literature which seeks to explain universal or recurring features of languages in terms of biases in learning (see Culbertson, 2012, forthcoming, for review): for instance, this approach has been adopted for morpheme order (St. Clair et al., 2009; Saldana et al., 2019), word order (Culbertson et al., 2012; Culbertson & Adger, 2014), correlations between word order flexibility and casemarking (Fedzechkina et al., 2017; Fedzechkina & Jaeger, 2020), asymetries in number marking (Kurumada & Grimm, 2019), preferences for certain types of phonological patterns (White, 2014; Martin & White, 2019), or general preferences for regularity or predictability of variation (e.g. Hudson Kam & Newport, 2005, 2009; Smith & Wonnacott, 2010). A common technique is to train participants on artificial linguistic systems in order to test whether configurations found more commonly among the world's languages are learnt more quickly or more accurately (e.g. Wagner et al., 2019), whether generalisations in the absence of direct evidence favour cross-linguistically more common configurations (e.g. Culbertson & Adger, 2014), or whether errors in learning (i.e. deviations from the input) tend to be skewed towards more common patterns (e.g. Culbertson et al., 2012).

FNJ use the latter type of design to test whether there is a bias in learning favouring Differential Object Marking (their Experiment 1) and Differential Subject Marking (their Experiment 2). Their participants were trained, over 4 days, on an artificial language for describing events involving interactions between two entities. In Experiment 1 events involved an animate subject and an animate or inanimate object (e.g. a chef hugging a referee, a mountie punching a chair); in Experiment 2 events involved an animate or inanimate subject and an inanimate object (e.g. a chef pushing a tricycle, a car dragging a shopping cart). In both experiments the input language participants were trained on featured variable word order and variable casemarking. Participants encountered both SOV (Subject-Object-Verb) and OSV (Object-Subject-Verb) orders during training (60% SOV, 40% OSV), and a case marker (a suffixal ending on the noun) occurred on 60% of objects (Experiment 1) or subjects (Experiment 2). Importantly, in the training input the occurrence/non-occurrence of the case marker was not conditioned on the typicality of the argument being marked, i.e. in Experiment 1 inanimate (typical) and animate (atypical) objects both occurred with the case marker in 60% of trials. In other words, their participants were confronted with an input language where word order is not a reliable cue to argument roles, unmarked atypical arguments potentially introduce ambiguity (if a sentence in Experiment 1 features two unmarked nouns which refer to animate entities, which is the subject and which is the object?), and the language does not exploit differential marking in a targeted way to reduce ambiguity in such cases.

Participants were trained on the input languages using a mix of passive exposure (participants saw events and heard descriptions) and comprehension tests (participants were presented with a description and asked to click on the picture of the subject of the event described). At the end of each day's training (from day 2 onward) participants were prompted with events and asked to produce the appropriate description. FNJ were interested in whether participants would shift the input language towards a configuration where casemarking preferentially targeted atypical arguments (i.e. animate objects in Experiment 1, inanimate subjects in Experiment 2), as seen in Differential Case Marking systems in natural languages.

Both experiments produced data at least somewhat consistent with this hypothesis. In Experiment 1, participants were more likely to mark animate than inanimate objects on day 2, a Differential Object Marking configuration which persisted (with roughly the same magnitude) on days 3 and 4. In Experiment 2 participants showed a (non-significant) tendency to mark inanimate subjects *less* than animate subject on day 2 (i.e. an *anti*-Differential Subject Marking configuration); however, there was a significant change in this preference over days, and by day 4 atypical inanimate subjects were more frequently marked than animate subjects, although not significantly so. While the results of Experiment 1 seem much clearer, both experiments show some effects of argument typicality on casemarking, either a Differential Object Marking configuration which was present early and persisted across several days of training (Experiment 1) or a significant shift in behaviour towards a Differential Subject Marking configuration from days 2–4 (Experiment 2). Note that in both cases these results are consistent with either an efficient communication account (mark arguments only where ambiguity would arise in the absence of marking), or an iconicity explanation (mark atypical arguments), although FNJ interpret these results as indicating a bias in learning favouring communicative efficiency.

1.4. Conceptual and methodological concerns

As we mention above, there is a growing body of literature using artificial language learning experiments to demonstrate biases in learning that favour common features of languages, or to provide learning-based accounts of skewed typological distributions. While FNJ sits within this tradition, it is slightly unusual in suggesting that biases in learning favour language features which increase communicative utility (although it is not unique in this regard: see also e.g. Carstensen et al., 2015; Fedzechkina et al., 2017; Levshina, 2018; Roberts & Fedzechkina, 2018; Kurumada & Grimm, 2019; Fedzechkina & Jaeger, 2020, which we return to in the discussion, or indeed Smith, 2004 for a similar line of argument based on evolutionary modelling). Most other accounts of biases in learning are motivated by arguments from naturalness (phonological systems which mirror perceptual or articulatory naturalness are easier to learn, Martin & White, 2019), representational simplicity (i.e. simpler patterns are easier to learn and more likely to be inferred, e.g. Moreton & Pater, 2012; Culbertson & Kirby, 2016), more general expectations of regularity/predictability (e.g. Hudson Kam & Newport, 2005, 2009; Reali & Griffiths, 2009; Smith & Wonnacott, 2010; Smith et al., 2017), or from preferences for iconicity in the meaning-form mapping, either at the lexical level (where signals resemble aspects of their meaning, e.g. Imai et al., 2008; Thompson et al., 2012, see e.g. Dingemanse et al., 2015; Nielsen & Dingemanse, 2020 for review), or at the structural level, where the structure of signals mirrors the structure of semantics (Schouwstra & de Swart, 2014; Culbertson & Adger, 2014)[5]. On the basis of these findings, we have argued that biases in learning may be at best agnostic to communicative function and often work against communicative utility, favouring simplicity even if that comes at a cost to communication. For instance, Kirby et al. (2008) and Silvey et al. (2015) show that artificial languages which are repeatedly learned and reproduced in an iterated learning paradigm (where the output from one learner becomes the input to the next learner in a chain of transmission) rapidly simplify and therefore lose the ability to encode distinctions.[6] Instead, Kirby et al. (2015) argue that communicative pressures only operate during communication—i.e. interaction between individuals with the goal of exchanging information by encoding semantic distinctions linguistically. From this perspective, languages must satisfy pressures for both learnability (imposed during learning) and communicative utility (imposed during communication), where these pressures are often in competition, for example where preferences for simplicity in learning conflict with pressures for expressivity in communication (see Carr et al., 2017; Kanwal et al., 2017a,b; Carr et al., 2020; Smith et al.,

[5]A noteworthy feature of both Schouwstra & de Swart (2014) and Culbertson & Adger (2014) is that these biases play out in generalisation or pure improvisation, rather than as an advantage in speed or accuracy of learning per se. In Schouwstra & de Swart (2014) participants are asked to improvise gestures to convey events and do so in a way where the order of gesture components reflects properties of the event being described; in Culbertson & Adger (2014) participants are trained on simple phrases but asked to generalise to the structure of more complex phrases. These experiments therefore provide evidence of prior biases with respect to structural iconicity, but through a slightly different means than used in the other learning studies reviewed above. Studies showing a straightforward learnability advantage for structural iconicity appear not to exist in the published literature.

[6]This is the case for paradigms where participants are trained via passive exposure (Kirby et al., 2008) or when training is framed in such a way that it emphasises the utility of encoding distinctions (Silvey et al., 2015); recall that the FNJ training method combines both these training methods

2020 for experimental treatments, and Regier et al., 2007; Kemp & Regier, 2012; Hahn et al., 2020 for evidence of this trade-off in a number of otherwise quite distinct domains).

In contrast, FNJ's results suggest that biases in learning might be the whole story (or at least a major part of it). Under their view, even aspects of language design which look like adaptations for communication are in fact a reflection of biases in learning. This would be important for at least two reasons. Firstly, if found to be generally applicable, it would offer an avenue to simplify theories of recurring features of language design. In particular, language design would then reflect biases in learning alone, rather than a compromise between partially-conflicting biases in learning and use. Second, it would pose a set of important new questions regarding the source of those biases in learning.[7]

Setting aside these conceptual issues about the relative role of learning and communication in shaping linguistic systems, there are also some unusual features of FNJ's method and findings which make replicating their results desirable. First, the two experiments show slightly inconsistent results in terms of how participants used case markers over days 2–4. In their Experiment 2, participants are on average (i.e. collapsing over participants and over argument animacy) already at day 2 producing case markers at roughly the frequency they encountered in their training data, i.e. on 60% of trials; this stays roughly constant across all 4 days. This is the result we would typically expect in a frequency-learning experiment, and is often known as probability matching (see e.g. Hudson Kam & Newport, 2005; Vouloumanos, 2008; Hudson Kam & Newport, 2009; Ferdinand et al., 2019, although it is important to note there are studies where participants diverge reliably from their input frequency, including when the input is complex, Hudson Kam & Newport, 2009 or one of the competing variants is dispreferred on other grounds, Culbertson et al., 2012). However, in FNJ's Experiment 1 participants quite strongly undershoot on their production of casemarking, producing roughly 25% of casemarked objects after two days of training (320 presentations of training sentences); they converge over days 3–4 on the 60% level in their input, but retain the over-marking of animate objects present from day 2. While this kind of fluctuation is entirely possible in two $N = 20$ experiments, the mismatch with the more accurate frequency tracking seen in Experiment 2 raises the possibility that something unusual happened in Experiment 1, which is potentially important since it is Experiment 1 which produces the more convincing Differential Case Marking effect.

Second, as FNJ directly acknowledge, their Experiment 2 produces striking different (and much weaker) pattern of results than Experiment 1: while there is an effect of animacy on casemarking, this plays out in an interaction with day, such that the effect of animacy on casemarking is different at day 4 than day 2: however, there is no significant effect of animacy on casemarking at either day 2 or day 4, and therefore no direct evidence of a Differential Subject Marking configuration anywhere. FNJ attribute this to a weaker bias

[7]For instance: are they an acquired bias that adult participants bring to the experimental task, based on their experience with well-designed natural languages? Or are they based on other expectations that languages should be structured to allow efficient communication, either arising from non-linguistic experience or innate biases about the expected form of linguistic systems? In the latter case, we would ultimately have to provide an evolutionary explanation of how such biases in learning could evolve, and in particular reconcile those results with evolutionary models which suggest that strong, domain-specific biases in language learning should be hard to evolve (Thompson et al., 2016).

for Differential Subject Marking; a more prosaic explanation is that these effects are at least somewhat fragile, making additional confirmation worthwhile.

Third, their stimuli are configured in a way that seems likely to inhibit, rather than encourage, Differential Case Marking. Their events (the events participants see labelled during training and are required to label at test, e.g. an animation of a mountie hugging a chef) are configured so that unmarked objects (Experiment 1) or subjects (Experiment 2) should *not* in practice be ambiguous. For example, in Experiment 1, participants encounter 10 animate referents (including a mountie and a chef) but 5 of these only ever occur as subjects and 5 only ever occur as objects (e.g. the mountie is only ever a subject, the chef is only ever an object). FNJ report that their participants were not sensitive to this regularity: when hearing a description involving the mountie and the chef, with no casemarking on the chef as object, participants were equally likely to select the mountie and the chef when prompted to select the subject, despite the fact that they had only ever seen the chef as an object, never a subject. This suggests that descriptions featuring unmarked animate objects were genuinely ambiguous for their participants. We found this surprising, based on our subjective impression when developing an experiment based on FNJ's Experiment 1, where we felt this regularity was highly salient. Indeed in a pilot experiment modelled on FNJ's Experiment 1, our participants were above chance on unmarked animate objects even on day 1, with performance increasing over days until they were near-ceiling on day 4, indicating that they rapidly identified the fact that some animate referents were never subjects and were able to use this to disambiguate in the absence of case markers. This potentially undermines FNJ's conclusion that biases in learning favour Differential Case Marking because of its efficiency advantages, since the efficiency of differential marking relies on ambiguity; in the absence of ambiguity, the efficient solution is to *never* mark case.

There are therefore three interesting avenues to explore in relation to this aspect of their method: 1) it may be that the (theoretical, and in our pilot data actual) lack of ambiguity of unmarked arguments in these stimuli results in FNJ *underestimating* the strength of biases in learning favouring Differential Case Marking—in stimuli where unmarked arguments are maximally ambiguous (e.g. where all animate objects can also appear as subjects), a learning bias for differential marking may be clearer; 2) relatedly, if differential marking in these experiments is dependent on participants' perceptions of ambiguity, their differential marking effects might disappear for participants who were able to disambiguate based on their knowledge of the restricted set of events they encountered in the experiment; 3) if their differential marking results replicate regardless of the ambiguity of unmarked arguments, this might suggest that the biases seen in FNJ are not to do with disambiguation per se, but rather with iconicity, a preference to mark atypical arguments. This latter possibility, if borne out, would bring their results into line with the experiments reviewed above on iconicity biases seen elsewhere in artificial language learning, and the more general claim that the pressures operating in learning are distinct from those operating in use (while potentially sometimes favouring similar outcomes Carr et al., 2020). Untangling this issue therefore potentially speaks to the debate over whether differential case marking has its roots in ambiguity avoidance or iconicity.

1.5. Our experiments

We report a series of 3 experiments where we set out to address these issues. In Experiment 1 we run a conceptual replication of FNJ's Experiment 1, targeting Differential Object Marking (henceforth DOM), the tendency to mark atypical [animate] objects. We make two changes to the design. First, we manipulate event structure: for some participants unmarked animate objects will be genuinely ambiguous. Second, we add a communicative test: rather than simply producing descriptions for a series of events in a non-communicative recall test, as in FNJ, participants use the language to communicate with a simulated interlocutor. Our Experiment 1 produces a rather different pattern of results to FNJ Experiment 1, more closely resembling the results of their Experiment 2. Initially we see participants marking *typical* objects more often than atypical objects (i.e. an *anti*-DOM configuration), but this declines over days until animate and inanimate objects are equally likely to be marked on day 4. Experiment 1 also suggests no influence of the presence or absence of ambiguity, which is problematic for the efficient communication account. Finally, Experiment 1 also shows that participants rapidly switch to a DOM configuration in communicative interaction where ambiguity becomes relevant.

In Experiments 2 and 3 we manipulate the statistics of participants' training data in order to try to understand the pattern of results in Experiment 1. What we find suggests that the behaviour we see in Experiment 1 reflects an iconicity bias early in learning, where atypical objects receive atypical marking, even if that is not the communicatively optimal configuration. This early bias is gradually overcome by additional training on the input language (which, by design, provides evidence against DOM). While we robustly find no evidence of DOM on day 4 in learning, DOM emerges rapidly in communicative interaction, replicating our results from Experiment 1. Additionally, the timecourse of the emergence of DOM in interaction is influenced by participants' experience of ambiguity, both for themselves and their communicative partner; participants' use of case markers is responsive to the fine-grained details of the communicative task they face. In sum, these experiments suggest that there is no bias in favour of communicatively-efficient Differential Object Marking in learning. Rather, communicatively-efficient configurations emerge in actual communication. We argue that the biases seen early in learning in these experiments in fact relate to iconicity (a bias for isomorphism between events and their linguistic expression), modulated by statistical features of the input.

2. Experiment 1

In Experiment 1 we replicate FNJ's Experiment 1, exploring the emergence of DOM during learning, and add a manipulation of event composition. For half of our participants, stimuli are structured as per FNJ—unmarked animate objects are potentially not ambiguous. The rest of our participants encounter events where each animate entity can act as both subject and object, and unmarked animate objects are therefore ambiguous — this should amplify the DOM effect seen by FNJ if it is driven by communicative efficiency.

2.1. Method

2.1.1. Participants

Participants were recruited via Amazon Mechanical Turk (MTurk), and were self-reported native speakers of English aged 18 or over. The 4 days of the experiment were launched as 4 separate HITs (Human Intelligence Tasks) on consecutive days. The day 1 HIT was open to anyone who possessed the MTurk qualification indicating they were based in the US; access to subsequent days was controlled with the use of MTurk's qualification system, with participants only qualifying for day 2 after completing day 1 and satisfying our progression requirements (see below), and so on for subsequent days. Participants were paid between $4 and $8 for each day (see below). The total number of participants participating on each day was: Day 1: 145; Day 2: 122; Day 3: 112; Day 4: 100.[8]

2.1.2. Stimuli and target language

As in Experiment 1 from FNJ, participants were exposed to an artificial language for describing events in which animate subjects acted on animate or inanimate objects. The grammar of the target language is given in Figure 1. Word order in the target language was variable, with both Subject-Object-Verb (SOV, where the subject noun preceded the object noun) and Object-Subject-Verb (OSV) order occurring; the target language uses SOV more than OSV (60% SOV, 40% OSV). Case was optionally marked on objects (by the addition of a suffix), and objects are casemarked more frequently than not (60% of objects are casemarked). Finally, as in FNJ, word order and casemarking were not independent; SOV sentences were more likely to feature casemarking, a point which we return to in the discussion of Experiment 1 and in Experiments 2–3: two thirds of SOV sentences are casemarked, whereas only half of OSV sentences are casemarked.

Importantly, training sentences were constructed such that animacy of the object did not condition word order or casemarking: the frequency of SOV word order and casemarking were the same for descriptions of events involving animate and inanimate objects, and there was therefore no DOM-like over-marking of animate objects in the participants' input. There was also no lexical conditioning: word order and casemarking was not conditioned on the identity of the nouns, verbs, or their combination.

Language stimuli were presented both as text and aurally: sound files were generated using the Tessa voice in the MacTalk speech synthesizer, with pitch and tempo increased by 30% using Audacity to produce a more monstrous/comical effect befitting the monster tutor who trained participants on the target language.

Rather than using animations as in FNJ, we used drawings to depict events.[9] These drawings included a set of 15 referents (10 animates: artist, boxer, burglar, chef, clown,

[8] Of these participants, 1 failed the noun comprehension test (see below) on day 1, and does not feature in any of the analyses that follow. 1 failed the sentence comprehension test (see below) on day 1; we included this participant in the analyses of identification accuracy, but this participant did not complete sentence production trials and therefore did not contribute data to any of the analyses on casemarking or word order. Note that the number of participants declines over days as not all participants return for all 4 days.

[9] These drawings were based on the stimuli from Branigan et al. (2000), which we adapted for use in a pilot study. Many thanks to Sara Rolando for producing these drawings, to Hanna Jarvinen for assisting in their preparation, and to Holly Branigan for providing the original stimuli.

$S \rightarrow N_{subject}\ N_{object}\text{-}ka\ V$ (SOV, casemarked, $p = 0.4$)

$S \rightarrow N_{subject}\ N_{object}\ V$ (SOV, not marked, $p = 0.2$)

$S \rightarrow N_{object}\text{-}ka\ N_{subject}\ V$ (OSV, casemarked, $p = 0.2$)

$S \rightarrow N_{object}\ N_{subject}\ V$ (OSV, not marked, $p = 0.2$)

$N \rightarrow \{$*slagum, tombat, nagid, melnog, norg,*
daf, plid, klamen, dacin, zub,
vams, bliffen, rungmat, lombur, groost$\}$

$V \rightarrow \{$*slergin, prog, shen, zamper*$\}$

Figure 1: The grammar for Experiment 1. Annotations give the proportion of the sentences in the target language exhibiting each order. There are four possible sentence types, exhibiting two word orders (SOV and OSV) and the presence or absence of a suffix (-ka) on the object. There is an overall preference for SOV order and casemarking (both of which occur on 60% of sentences); as per FNJ, casemarking and word order are correlated, such that SOV sentences are more likely to be casemarked than not, whereas OSV sentences are as likely to be casemarked as not. The target language consists of up to 15 nouns and 4 verbs (see main text); these labels were assigned to referents/actions at random for each participant (e.g. *slagum* might refer to the police officer for one participant and the medic for another).

cowboy, dancer, medic, police officer, waiter; 5 inanimates: ball, cake, cup, jug, top hat) and 4 actions (kicking, punching, shooting and touching/pointing). Note that we use fewer actions than FNJ, who had 8 distinct actions.

2.1.3. Manipulation of event composition

As discussed above, in FNJ the set of events were constructed such that subjects were always animate, objects were equally likely to be animate or inanimate, but the set of animate subjects and animate objects were distinct. In other words, animate objects were never animate subjects and vice versa (e.g. if the waiter appeared as a subject, she would never appear as an object). We manipulated event composition between participants. We ran one condition following the FNJ scheme, which we will refer to as the Subjects Cannot Be Objects condition — for each participant 5 animates were randomly assigned to act as subjects and 5 different animates as objects. We also ran a second condition where the same set of 5 animate referents appeared equally often as subjects and as objects, which we refer to as the Subjects Can Be Objects condition; the set of animate referents used was selected randomly for each participant, and events were constructed such that the subject and object in any given event were distinct (i.e. participants never saw the event where the waiter punched the waiter). In order to match the experiments in terms of frequency of events involving animate and inanimate objects, this necessitated a difference between conditions in the total number of nouns in the target language: in the Subjects Cannot Be Objects condition each participant learnt a miniature language featuring 15 nouns (5 for animate subjects, 5 for animate objects, 5 for inanimates), whereas in the Subjects Can Be Objects

condition each participant learnt 10 nouns (5 for animates, which appeared as both subjects and objects, 5 for inanimates).

2.1.4. Procedure

The experiment was coded in Javascript, and participants completed the experiment in a web browser. Participants were briefed that they would learn the language spoken by a monster named Smeeble; its language was called Smeespeak. We based our procedure closely on that of FNJ, with modifications intended to keep the overall experiment duration and difficulty manageable, and to retain the attention of on-line participants; we also added an interaction phase at the end of day 4. A full session consisted of the 7 phases detailed below: on day 1, participants only completed the first 5 of these, with sentence testing beginning only on day 2; on day 4 participants additionally completed an interaction phase.

Noun training 1 (all days): On each noun training trial, participants were shown an animate or inanimate referent and heard its name in Smeespeak. Two buttons were shown onscreen, showing the object's correct label and another randomly-selected noun. Participants were instructed to click on the word that matched what they heard. Clicking on a label resulted in that word appearing below the referent object; if the participant clicked on the correct label the label appeared in green and they progressed to the next trial; if they clicked on the wrong label it appeared in red and they repeated the trial. Participants received either 20 or 30 such trials (depending on event composition), presenting each of the nouns twice in random order.

Noun comprehension test 1 (all days): On each noun comprehension trial, participants saw two referents on-screen (one target and a randomly-selected foil drawn from the set of 10 or 15 referents, depending on event composition), heard the name of the target being spoken by Smeeble, and saw the target noun on-screen as text. They were instructed to click on the object that matched what they heard. Correct responses resulted in a success sound, a happy Smeeble and the addition of 10 points to their running total; incorrect responses resulted in a failure sound, a sad Smeeble, and no points; participants progressed to the next trial regardless of their success or failure. Participants received 10 or 15 such trials, presenting each of the nouns once in random order.

Sentence training (all days): See Figure 2(a-b) for examples. Prior to beginning sentence training, participants were prompted to "pay close attention to the words AND the order they are in, so that later on you can describe things for yourself." On each sentence training trial, participants saw an image of an event (e.g. a medic shooting a hat; a dancer shooting a medic), heard its 3-word description in Smeespeak, and saw the description as text on-screen. Under the text description were three buttons, labelled with the subject noun, uninflected object noun, and verb from the sentence they had just heard; the buttons appeared in the same order as the description they heard/saw (e.g. for an SOV sentence, the buttons appeared in the order subject, object, verb). On each trial participants were either prompted to "click on the one DOING the action", "click on the one the action is DONE TO", or "click on the ACTION". The prompt was randomly selected at each trial. If the participant selected the correct

button (e.g. clicking the subject noun when prompted to do so) then the relevant word in the text was briefly highlighted in green; if they clicked on the wrong button then the word they clicked was highlighted in red and they repeated the trial. Participants received 80 such trials, featuring each action 20 times, each possible subject 16 times (4 times with each verb), and each possible object 8 times (twice with each verb).

Noun training 2 (all days): This worked in exactly the same way as noun training 1, but participants were warned that the training phase would be followed by a test which had to be passed to allow progression further into the experiment.

Noun comprehension test 2 (all days): This worked in exactly the same way as noun comprehension test 1, but participants were instructed that progression to the next stage of the experiment depended on satisfactory performance. Participants who scored below 70% correct on this test (i.e. selected the incorrect referent on 4 or more of 10 trials in the Subjects Can Be Objects condition, or 5 or more of 15 trials in the Subject Cannot be Objects condition) did not progress to the next stage of the experiment, did not qualify for the next day, and were paid $4 for their participation. We adopted the 70% threshold from FNJ.

Sentence comprehension test (all days): See Figure 2(c-e) for example trials. Prior to beginning this phase, participants were again warned that progression depended on satisfactory performance. Sentence comprehension worked in a similar way to the noun comprehension phases, in that participants were prompted with a description (presented aurally and in text) and had to click on an image in response from an array of two choices. Rather than hearing individual nouns as in noun comprehension trials, participants heard and saw 3-word sentences; the two objects they selected between were the two referents mentioned in the sentence (i.e. an animate and an inanimate, or two animates), and their instruction was always to "click on the one DOING the action". As in noun comprehension trials, participants received feedback on their response (happy or sad Smeeble, sounds, points for correct responses), and progressed to the next trial regardless of their accuracy. Participants received 80 such trials, constructed in the same way as in sentence training. Participants who scored below 70% correct during sentence comprehension did not progress to the next stage of the experiment, did not qualify for the next day, and were paid $6. On day 1, this was the final phase, and participants were paid $6 and qualified for the next day if they met our progression criterion.

Sentence production test (days 2+ only): See Figure 2(f) for an example trial. On each trial, participants were shown a picture depicting an event (e.g. a dancer shooting a hat) and were asked to provide the appropriate description in Smeespeak by typing into a text box. Participants were provided with the appropriate verb for each trial; each trial therefore involved recalling the appropriate nouns, generating a word order and deciding whether to include the case marker. Participants received 40 such trials, featuring each action 10 times, each possible subject 8 times (twice with each verb), and each possible object 4 times (once with each verb). On days 2 and 3 this was the

final phase of the experiment, and participants reaching this point were paid $6 and qualified for the next day.

Interaction (day 4 only): Participants played a director-matcher game in which they alternated describing an event for Smeeble, and selecting an event based on Smeeble's description. Director trials resembled sentence production trials (Figure 2(f)); Figure 2(g) gives an example matcher trial. When directing, participants were presented with an event and prompted to type the description so that Smeeble could identify it. On matcher trials, they were presented with a description from Smeeble and two events, and asked to select the event which Smeeble was describing. The pair of events the matcher had to choose between contained the target event and a foil; the foil was selected such that encoding the identity of the subject and object was required for reliably successful communication.[10] Each successful interaction was rewarded as in earlier phases, i.e. with a happy or sad Smeeble, a sound, and points. The success of the interaction when the participant was matcher was determined by their response, i.e. whether they clicked the target picture. When the participants was the director and Smeeble was the matcher, the experiment software simulated a rational matching behaviour based on the same matcher task as faced by the participant,[11] and feedback was based on whether that resulted in Smeeble selecting the correct event or not. Note that, when acting as matcher, Smeeble did not 'know' that some objects are never objects in the Subjects Cannot Be Objects condition, and therefore events involving two animates required casemarking to be reliably interpreted correctly by Smeeble in both event compositions. The participant was equally likely to be director or matcher on the first interaction trial, and roles alternated thereafter; participants acted as director for 40 trials and as matcher for 40 trials, with the sets of events and Smeeble's

[10]The foil was of one of two types, each occurring with equal probability. If the target event object was inanimate, then on half of trials the foil event had the same subject and action as the target event, but a different object (selected at random from the set of possible objects, excluding the object in the target event) and on half of trials the foil had the same verb and object as the target, but a different subject (selected at random from the set of possible subjects, excluding the subject in the target event). If the target event object was animate, on half of trials the foil had the same verb and object as the target, but a different subject (selected randomly as above) and on half of trials the foil event had the same subject, verb and object as the target, simply with roles reversed (i.e. the subject in the target was the object in the foil).

[11]On each trial where Smeeble was matcher, the software generated a matcher array composed in the same way as the participant's matcher array, i.e. consisting of the target event and another similar foil event. Each of those events has 4 grammatical descriptions in Smeespeak (SOV or OSV order, object marked for case or not). If the participant's description exactly matched one of those descriptions then Smeeble selected that event; if the participant's description exactly matched the description for more than one event (as could occur if the object was not casemarked and the target event featured two animate objects) then Smeeble selected an event randomly; if the participant's description didn't exactly match any description then Smeeble selected the event whose description was closest to the description provided by the participant, where distance between descriptions was simply the number of words which were identical between the two descriptions; again, if multiple descriptions were equally close, Smeeble selected randomly between them. Note that this procedure is the same in both the Subjects Cannot Be Objects and Subjects Can Be Objects conditions, and therefore does not exploit the reduced ambiguity of events in the Subjects Cannot Be Objects condition - therefore, when both referents are animate, descriptions lacking casemarking are potentially ambiguous for Smeeble.

descriptions constructed as in the sentence training and sentence production phases. This was the final phase of the experiment on day 4; to compensate for the increased duration, participants progressing to this point were paid $8.

At the end of the day's session, participants were informed about their total payment, and given details of how to participate in the next day as appropriate.

2.1.5. Coding word order and casemarking

We automatically coded participants' word order and use of casemarking on sentence production and interaction trials. Each typed description was broken into words bounded by whitespace; for each word, we then identified the closest matching label from the trained vocabulary, allowing the possibility that the marker '-ka' could be appended to any word. We accepted as grammatical any word sequence which could be generated for the given scene by the target language (e.g. SOV or OSV order, object marked or unmarked) and excluded all other trials from the analyses that follow (e.g. where the noun used did not correspond to one of the referents in the event being described; where the case marker appeared on the subject or the verb; where another word order was used); this resulted in the removal of 26% of sentence test trials on day 2, falling to 17% on day 4; the vast majority of these rejected trials featured a single lexical error, i.e. using an incorrect noun for one of the two referents.[12] Note that all participants were included in our analyses; FNJ additionally excluded participants who always or never marked case (resulting in the exclusion of 5 participants in their Experiment 1), but we did not do this, since excluding participants who could not show the effect of interest (differential marking) from the analysis seemed anti-conservative to us.

All data for all the experiments reported here, plus graphing and analysis code used to generate all reported results, are available online at `https://github.com/kennysmithed/SmithCulbertsonDOM`.

2.2. Results

We begin by investigating whether our between-participant manipulation of event composition had any effects on participants' ability to interpret sentences correctly during sentence comprehension — in particular, did the manipulation of event composition lead to differences in the ambiguity of unmarked objects? If so, and if DOM is driven by ambiguity avoidance, then we might expect stronger DOM effects in the Subjects Can Be Objects condition, where we expect unmarked animate objects to be more ambiguous.

We then test whether, as in FNJ, participants preferentially case mark animate objects in non-communicative sentence recall (i.e. do we replicate the FNJ result?) and if so, whether this is influenced by event composition (again, predicting greater DOM in the Subjects Can Be Objects condition).

[12]FNJ included trials where one noun was 'mispronounced', on the basis that it is still possible to determine the relative order of the two nouns; however, this assumes that in such trials the other noun was produced as intended (and not, e.g., mapped to the wrong referent); making such case-by-case decisions in our larger sample would also be substantially more time-consuming. We therefore apply this stricter criterion.

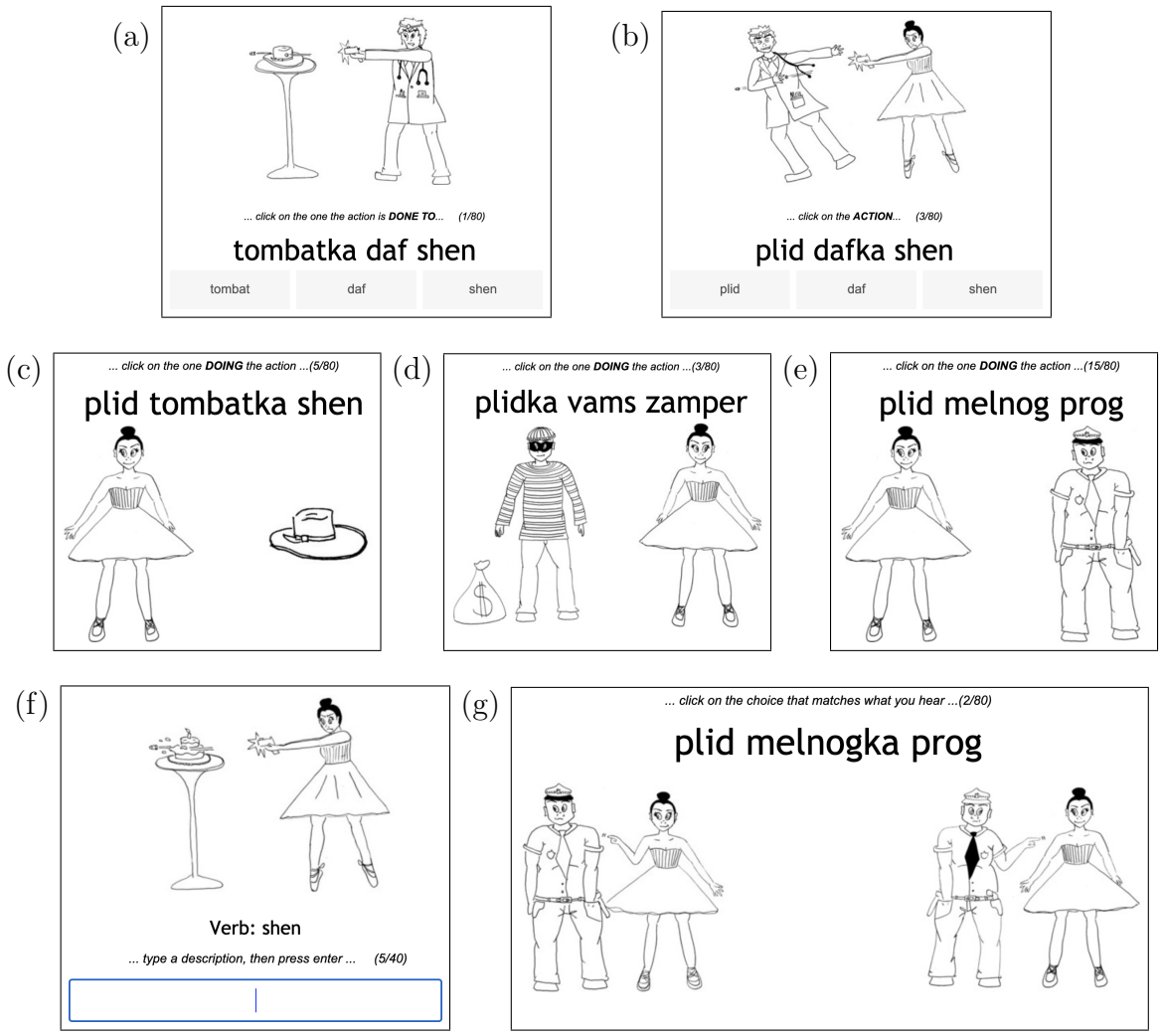

Figure 2: Example sentence training trials (a-b), sentence comprehension trials (c-e), a sentence production trial (f; this trial type also occurs during interaction) and a matcher trial during interaction (g). In training trials participants are required to select one word from the multi-word sentence. In comprehension trials participants are required to select the subject referent based on a provided description, including in trials with only one animate option (c), with two animates where the object (i.e. non-subject) is casemarked (d), and with two animates where neither is marked. In test trials, participants produce a description using a provided verb — these trials occur at the end of days 2–4 as a recall test, plus during interaction on day 4. In matcher trials during interaction, participants are prompted with a description from Smeeble and select an event. Note that these examples are drawn from the Subjects Can Be Objects condition, since some characters (*daf*, the medic) appear in both subject and object roles.

Third, we explore whether DOM-like effects appear in communicative interaction. Recall that this the only place where one would expect to see communicative efficiency considerations at play, if biases in learning (captured during the non-communicative recall tests) are agnostic with respect to communicative function. We again explore whether this is affected by event composition.

All analyses are conducted using in R Version 3.5.3 (R Core Team, 2019) using logistic mixed effects regression using the lme4 package version 1.1-21 (Bates et al., 2015).

2.2.1. Identification accuracy on comprehension trials

We begin by investigating whether participants could correctly identify the subject of sentences presented during sentence comprehension trials (where participants are presented with a sentence and on-screen images of the subject and object referent, and prompted to click on the subject), and whether this was affected by our between-participant manipulation of the composition of events. Recall that in the Subjects Cannot Be Objects condition the stimuli are constructed such that animate subjects are *never* objects, following FNJ; a participant who had noted this regularity would be able to correctly interpret sentences describing scenes involving two animates even if the object was not casemarked. In our Subjects Can Be Objects condition there was no such regularity in the set of events participants saw, which should reduce performance to chance levels for trials with an animate object and no casemarking (because either of the two animate referents the participant is required to choose between could be the subject).

Figure 3 shows the proportion of sentence comprehension trials on which participants were able to identify the subject. As is clear from the figure, participants in both conditions are unsurprisingly at ceiling performance from day 1 on trials featuring an inanimate object, virtually always correctly identifying the animate referent as the subject. Performance on trials with a casemarked animate object is similarly high in both conditions, even on day 1, indicating that participants rapidly learnt the disambiguating function of the case marker. In the Subjects Cannot Be Objects condition, in line with our subjective impression and pilot data, but in contrast to the result reported by FNJ, participants also performed well on trials where the object was animate but unmarked; by day 4, most participants answer *all* such trials correctly, indicating that the structure of the event set renders unmarked animate objects unambiguous. In contrast, in the Subjects Can Be Objects condition, performance on trials featuring an unmarked animate object was at chance (50% correct) throughout, as expected — the structure of the set of events in this condition renders such descriptions genuinely ambiguous.

These impressions are confirmed by a statistical analysis where we predict response accuracy (correct or incorrect) based on fixed effects of day, casemarking, event composition and their interactions (see Table 1 for details of contrasts coding, random effects, and full summary table) — we ran this model on data from trials featuring an animate object only, since these are the most relevant trials, and including animacy as a fixed effect led to convergence problems. Participants in the Subjects Cannot Be Objects condition perform well even on day 1 for unmarked animate objects: the odds ratio of selecting the subject correctly on these trials is estimated as roughly 1.6 to 1 (log odds of 0.49, $SE = 0.11$), which is significantly higher than would be expected under random guessing (where the odds ratio would be 1 and the log-odds 0; $p < .001$). This level of performance is roughly what would be expected if par-

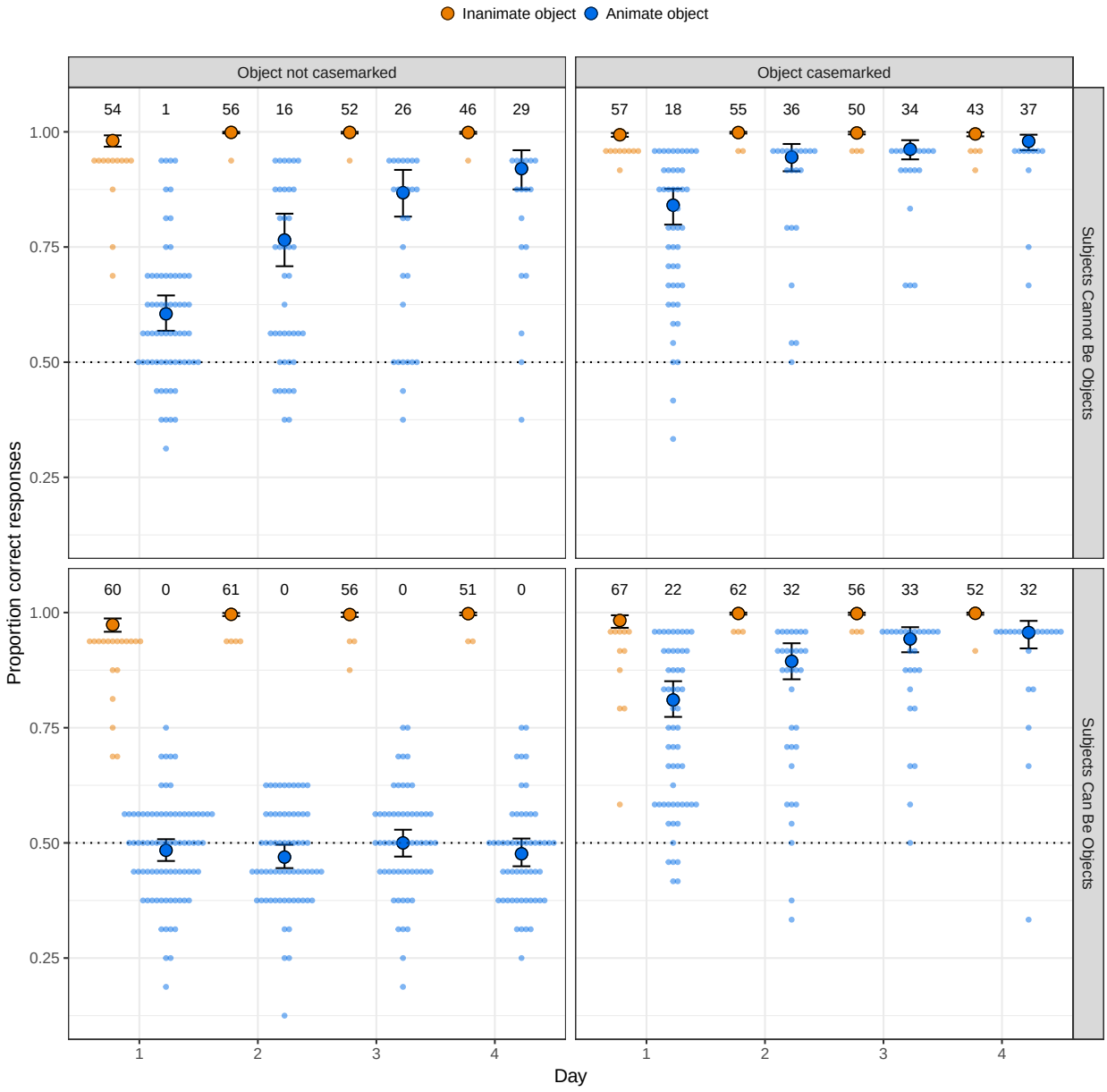

Figure 3: Proportion of correct responses on sentence comprehension trials in Experiment 2. Larger points give mean of by-participant means, error bars give bootstrapped 95% CIs on those means, dashed line indicates chance performance, dots give individual participant means. There are too many participants performing at ceiling (i.e. all correct responses) to plot; instead, the text annotations give the number of participants producing ceiling performance. Performance is at ceiling when the object is inanimate, and is high when the object is animate but casemarked; however, unmarked animate objects are only genuinely ambiguous in the Subjects Can Be Objects condition.

ticipants made optimal use of word order cues to agency (60% of sentences used SOV order, which would yield an expected log odds of success for an attentive learner of 0.41, which is not significantly different from the success rates of our participants; p = .412). However, there is no suggestion that participants in the Subjects Can Be Objects condition make use of the word order cue in this way, even on day 4, which suggests the above-chance performance in the Subjects Cannot Be Objects condition on day 1 derives from the structure present in the set of events; furthermore, performance on these unmarked trials continued to increase over the 4 days of training in the Subjects Cannot Be Objects condition (as indicated by significant effects of day, all ps $< .001$; day=2: $b = 0.85, SE = 0.11$; day=3: $b = 1.60, SE = 0.13$; day=4: $b = 2.24, SE = 0.16$), reaching levels on day 4 that cannot be attributed to the word order cue. Our data therefore provide strong evidence that participants in the Subjects Cannot Be Objects condition were aware from early on that some animate referents were less likely to be subjects or could not be subjects. Casemarking did however facilitate correct identification of the subject: this effect was present throughout (as indicated by a significant effect of casemarking on day 1, $b = 1.40, SE = 0.10, p < .001$, and no significant negative interactions at subsequent days) and particularly marked on day 2 (as indicated by a significant casemarking x day 2 interaction, $b = 0.40, SE = 0.18, p = .022$); the absence of positive interactions for days 3 and 4 presumably reflects the fact that performance on those days already approached ceiling even without casemarking.

The results for the Subjects Can Be Objects condition look rather different, as confirmed by multiple significant interactions involving event composition. Performance in the Subjects Can Be Objects condition is significantly lower on day 1 ($b = -0.57, SE = 0.14, p < .001$), with a log odds of correct responses of close to 0, consistent with chance performance; furthermore, the negative terms for the interactions between event composition and day are of the same magnitude as the positive terms for day in the Subjects Cannot Be Objects condition, indicating no day-by-day increase in performance on unmarked trials over days in the Subjects Can Be Objects condition. Finally, while the Subjects Can Be Objects condition shows no difference from the Subjects Cannot Be Objects condition in the effect of casemarking at days 1 and 2 (as indicated by n.s. effects), at days 3 and 4 participants in the Subjects Can Be Objects condition continue to benefit from casemarking (as indicated by significant 3-way interactions involving day 3 and day 4; interaction at day=2: $b = 0.37, SE = 0.22, p = .098$; interaction at day=3: $b = 1.28, SE = 0.26, p < .001$; interaction day=4: $b = 1.63, SE = 0.32, p < .001$).

In sum, there is strong evidence that the unmarked animates were genuinely ambiguous in the Subjects Can Be Objects condition but not in the Subjects Cannot Be Objects condition, even on day 1. If the DOM effect seen by FNJ is due to participants compensating for the potential ambiguity of unmarked animate objects in their own productions, then based on these results we might reasonably expect stronger DOM in our Subjects Can Be Objects condition, where unmarked animates do genuinely introduce ambiguity.

2.2.2. Animacy and casemarking during sentence recall

Figure 4 shows the frequency of casemarking in the sentences participants produced during the sentence production test phase of the experiment, across days 2–4 of the experiment — recall that there was no sentence production test on day 1. Impressionistically, the results do not look like FNJ Experiment 1: numerically, atypical animate objects are marked *less*

on day 2, although this tendency seems to gradually reverse until by day 4 the configuration is more DOM-like, with animate objects being marked more. There is no obvious effect of event composition.

We ran a logit regression on this data, including day, animacy of object, event composition and their interactions as fixed effects (see Table 2 for the full set of results). Participants at day 2 produce casemarking at roughly the frequency seen in their input — the odds of producing a case marker are approximately 1.6 to 1 (log odds of 0.50, which is not significantly different from the log odds casemarking in the input, 0.41, p = .746). There is an effect of animacy at Day 2 ($b = -0.43, SE = 0.22, p = .047$), in the opposite direction to that seen in Differential Object Marking: animate objects are significantly *less* likely to be casemarked than inanimate objects. The effect of animacy changes over time, with participants' log-odds of casemarking animate objects being significantly higher at Day 4 than at Day 2 (as indicated by the significant positive interaction between animacy and Day = 4, $b = 0.69, SE = 0.23, p = .002$); note however that a model with the same fixed and random effects structure but setting day 4 as the intercept indicates no significant effect of animacy at day 4 ($b = 0.26, SE = 0.21, p = .218$), i.e. no DOM effect at day 4. This pattern of results (reversal from day 2 to day 4 but no significant effect of animacy on casemarking at day 4) is reminiscent of the unfolding Differential Subject Marking effect seen in FNJ Experiment 2.[13] Finally, and surprisingly given the substantial differences in the sentence comprehension trials, both the Subjects Cannot Be Objects and Subjects Can Be Objects conditions show comparable patterns of results, as indicated by the absence of significant effects of event composition or interactions involving event composition (smallest $p = .228$).

2.2.3. Animacy and casemarking during interaction

Participants underwent a final additional test on day 4, using the language they had learnt to interact with Smeeble, their monster language tutor. Figure 5 shows the proportion of successful communicative trials, split by the animacy of the object, the presence or absence of casemarking, and the participant's role on that trial. As in the analysis of sentence comprehension trials, unmarked animate objects are effectively unambiguous for participants in the Subjects Cannot Be Objects condition, but are ambiguous for participants in the Subjects Can Be Objects condition, although recall that in the role of matcher, Smeeble finds unmarked animate objects equally ambiguous in both.

Figure 6 shows the frequency with which participants casemarked objects in their productions during interaction, with their frequency of casemarking on the immediately-preceding day 4 sentence production test also plotted for comparison. There is an across-the-board increase in casemarking during interaction, with the largest increases seen for target events where the absence of casemarking is most problematic for communication, i.e. when the object is animate. However, there does not seem to be greater marking of animate objects in the Subjects Can Be Objects condition, where participants might be particularly conscious of the problems posed by unmarked animates. We ran a logit regression on participants' use of casemarking, predicting the probability of a case marker based on fixed effects of block (recall or interaction), animacy of object, event composition and their interaction (see

[13] We also obtained a similar pattern of results — an inverted DOM effect on day 2, disappearing by day 4, in a 4-day N=47 pilot experiment, using events similar to those in our Subjects Cannot Be Objects condition.

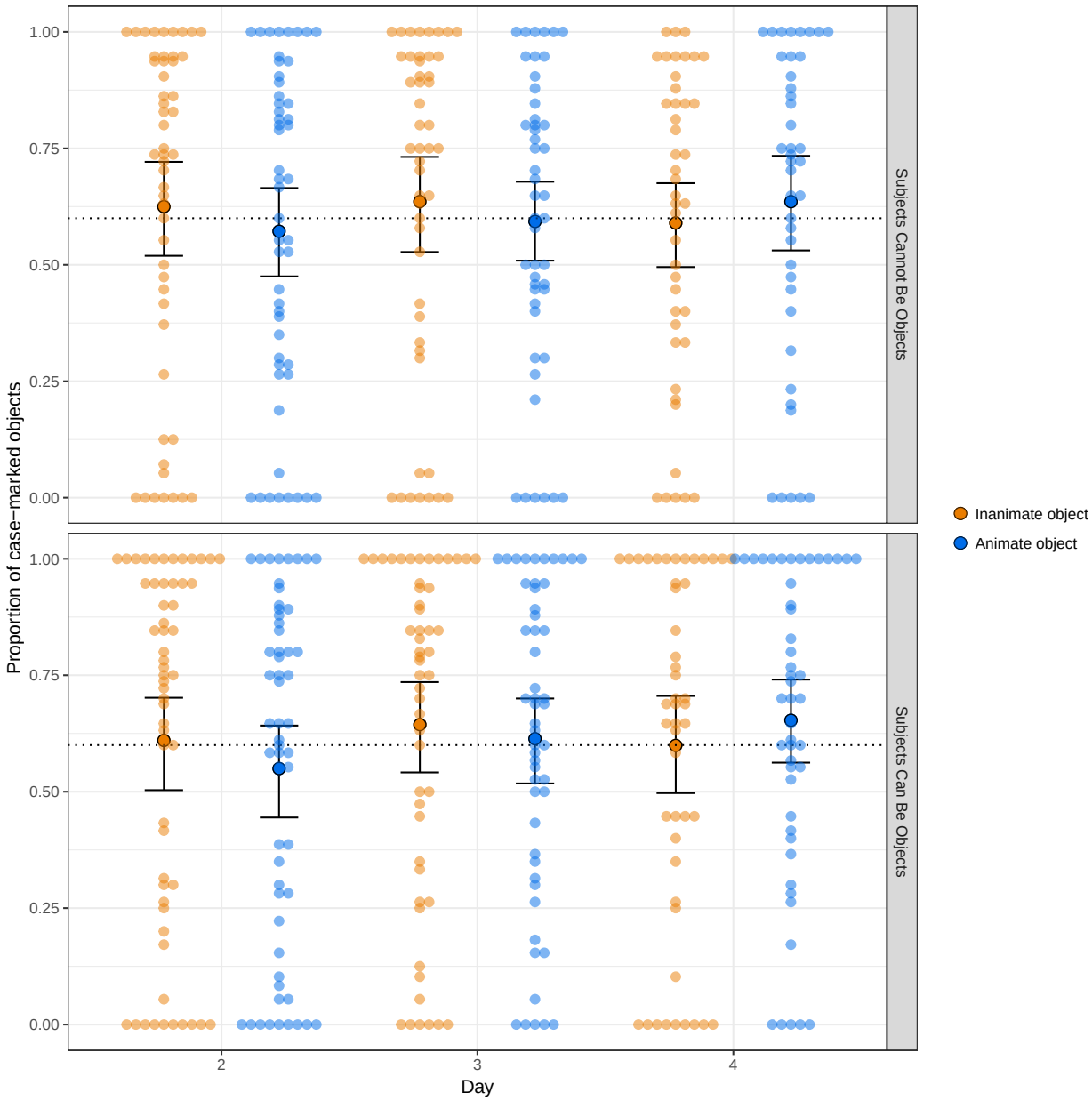

Figure 4: Proportion of casemarking in participants' productions in Experiment 1 during the sentence production test. Plotting conventions are as in Figure 3: Larger points give mean of by-participant means, error bars give bootstrapped 95% CIs on those means, dots give individual participant means. The horizontal dotted line shows casemarking frequency in the input (for both animate and inanimate nouns). Overall casemarking use approximates the input frequency. However, both conditions show an interaction between animacy and day: on day 2 there is a tendency to over-mark inanimate objects; by day 4 this tendency reverses, and participants case mark animate and inanimate objects at approximately the same frequency.

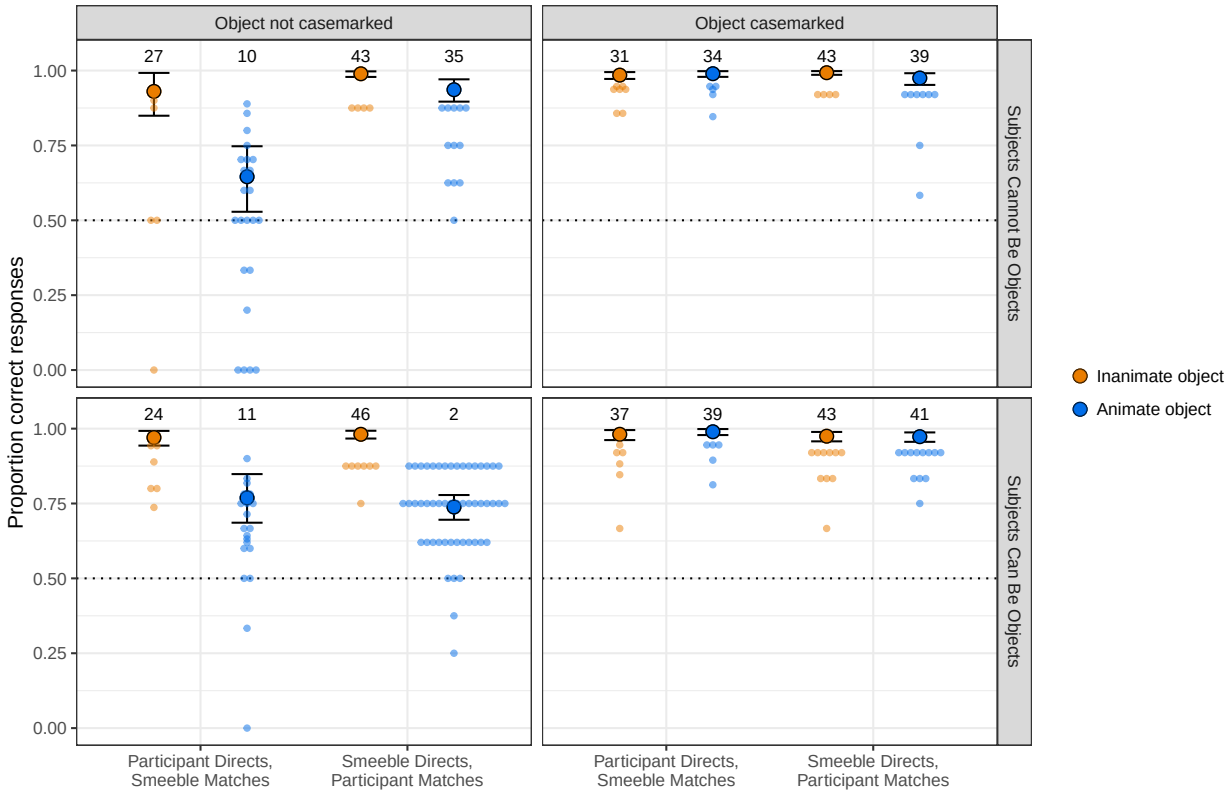

Figure 5: Proportion of successful trials during interaction with Smeeble on day 4. Plotting conventions are as in Figure 3 (NB. the annotations give number of participants performing at ceiling; these participants are not plotted individually). The horizontal dotted line shows chance performance on the two-choice task facing the matcher; note however, that when the object is animate, on half of trials the foil choice features an inanimate object, meaning that chance performance for a moderately attentive matcher (i.e. who realises that inanimates are never subjects) can achieve 75% accuracy even for unmarked animate objects. Mirroring the results for sentence comprehension, communicative accuracy is high when the object is inanimate or animate but casemarked; unmarked animate objects are always ambiguous when Smeeble is the matcher (performance averages around 75%), and for participants in the Subjects Can Be Objects condition, but are effectively unambiguous for participants in the Subjects Cannot Be Objects condition. Note that Smeeble's matching performance on unmarked trials has very high variance, because in some cases participants produced very few unmarked objects).

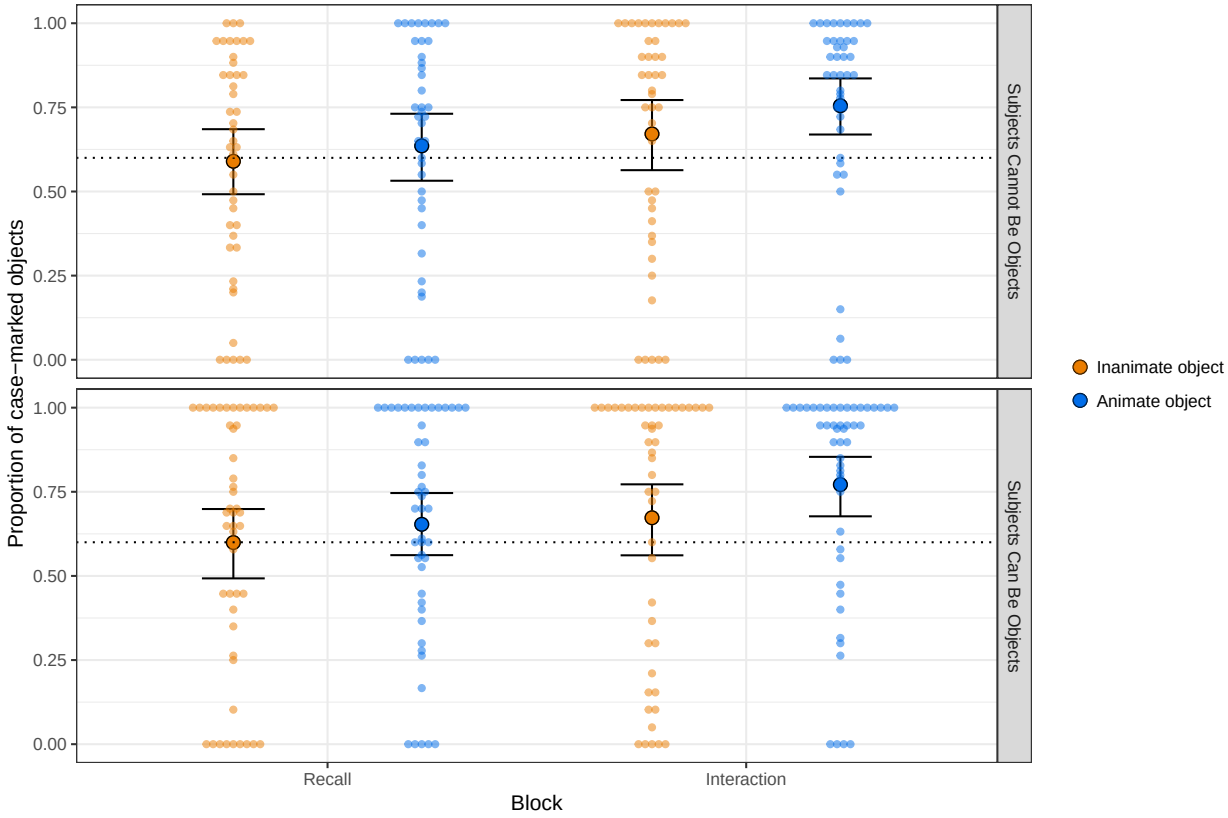

Figure 6: Frequency of casemarking in participants' productions on day 4 of Experiment 1, during the sentence production test (labelled 'Recall' here) and subsequent interaction with Smeeble. The horizontal dotted line shows casemarking frequency in the input (for both animate and inanimate nouns). Participants casemark more during interaction than the non-communicative recall test, and show an increased DOM-like over-marking of animate objects during interaction that is far less prominent (and statistically n.s.) in recall.

Table 3). This analysis indicated no effect of animacy during pre-interaction sentence recall ($b = 0.19, SE = 0.21, p = .363$), confirming that there was no differential marking of animate objects in the final sentence production test on day 4. There was however a substantial effect of block ($b = 0.95, SE = 0.22, p < .001$), indicating that participants casemarked more during interaction, and a more modest but significant interaction between block and animacy ($b = 0.43, SE = 0.20, p = .033$), indicating that this increase in marking was greater for animate objects, i.e. a DOM-like over-marking of animate objects developed rapidly in interaction. There were no significant interactions involving event composition (smallest $p = .417$), suggesting this effect was equivalent in both the Subjects Cannot Be Objects and Subjects Can Be Objects conditions. This may be because Smeeble's comprehension behavior—treating unmarked animates as ambiguous—is independent of our manipulation of event composition, a point we return to in several analyses in the remainder of the paper.

2.2.4. Animacy affects word order

In addition to the unfolding effects of animacy on casemarking, we noticed that animacy affected participants' choice of word order: as can be seen in Figure 7, participants are less

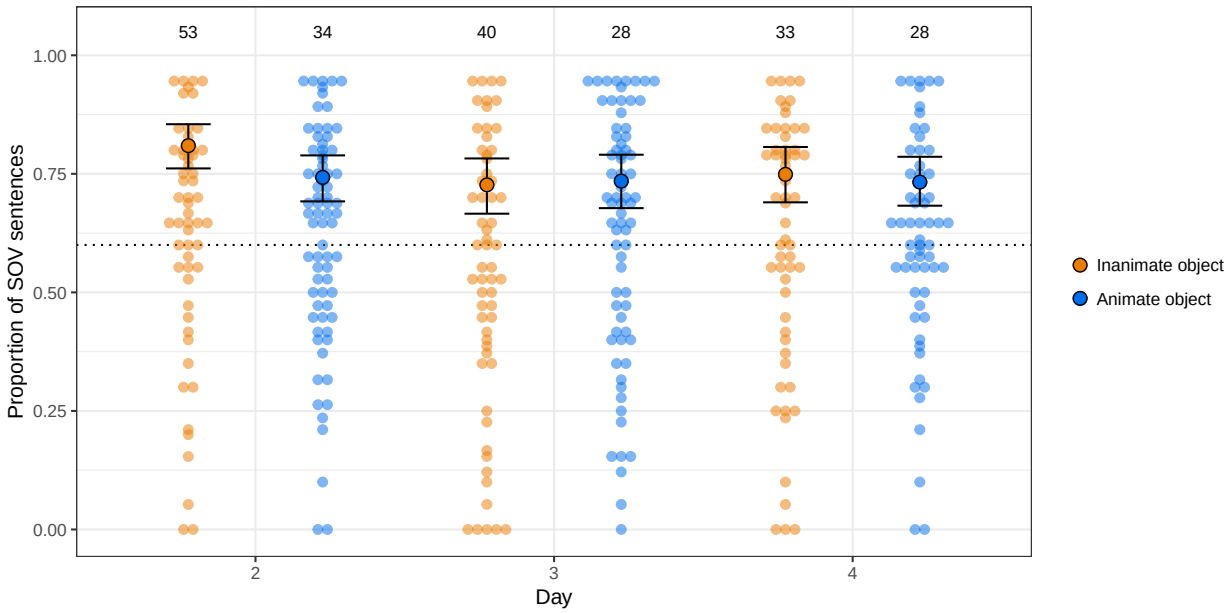

Figure 7: Proportion of SOV (as opposed to OSV) word order in participants' productions in Experiment 1 (collapsing across event compositions). There are too many participants at 100% SOV to plot, so those participants are indicated in the text annotations. The horizontal dotted line shows frequency of SOV in the input (for both animate and inanimate nouns). Participants over-produce SOV order across all 4 days, and show a preference for SOV order when the object is inanimate (or equivalently, a reduced preference for SOV over OSV when the object is animate), the magnitude of which declines at later days.

likely to use SOV order (and therefore more likely to use OSV order) when the object is animate, particularly on Day 2. As for our analysis of casemarking, we ran a logit regression on this word order data, predicting the probability of SOV order using a model with the same structure and coding scheme as used for the analysis of casemarking (fixed effects of day, animacy of object, event composition and their interaction; see Table 4). Participants have a substantial preference for SOV word order at Day 2 (the log-odds of producing SOV order are significantly higher than the frequency of SOV in the input, $p < .001$), which persists across days (as indicated by the absence of significant effects for Day). There was also a significant effect of animacy on word order choice at Day 2 ($b = -0.93$, $SE = 0.24$, $p < .001$), indicating that participants were less likely to use SOV order (i.e. more likely to use OSV order) when the object was animate; this conditioning of word order on animacy diminishes in subsequent days (as indicated by significant interactions between animacy and day at days 3 and 4: Day=3, $b = 0.55$, $SE = 0.14$, $p < .001$; Day=4, $b = 0.33$, $SE = 0.15$, $p = .028$). There was no evidence for any effects of event composition on word order choice.

This reduced preference for SOV and increased use of OSV when the object is animate is likely due to the relative salience of animate and inanimate referents. There is a widely-observed tendency to produce animate or human entities before inanimate or non-human entities (e.g. Meir et al., 2017). In events with an animate subject and an inanimate object, this preference to order animates before inanimates leads to the use of SOV order; in events with two animate entities (i.e. when the object is animate), the two referents are

better matched on this dimension and therefore there is a reduced preference for SOV order and increased use of OSV order. While this observed effect therefore meshes with an independently-motivated bias in production, its presence here potentially complicates the interpretation of the changing effects of animacy on word order that we see develop over the 4 days of our experiment, as discussed further below.

2.3. Discussion

Experiment 1 produces results which only partially support those from FNJ. We do find an effect of animacy on casemarking, and that effect changes over the course of the 4-day experiment, as the language shifts towards a DOM-like configuration where the tendency to mark animates is higher on day 4 than it was on day 2. However, the pattern of results is far weaker than those from FNJ Experiment 1: where they found a DOM-like effect from day 2, we see a significant *inverted* DOM configuration on day 2, with the interaction between day and animacy being largely driven by the elimination of this inversion; on day 4 during sentence recall the effect of animacy on casemarking is not significant. This pattern of results: an initially-inverted effect, which disappears on day 4, is closer to the results from FNJ Experiment 2, for Differential Subject Marking. Under a generous interpretation, it therefore provides further evidence that learners show a tendency, developing over days, to change their use of case markers in ways which are consistent with a bias in learning favouring efficient communication.

However, several aspects of our data are problematic for the efficient-communication-in-learning account. First, it's not clear why the effect of animacy would unfold over multiple days — most models of learning would predict that the effect of prior biases are most pronounced when the learner has relatively little data, and are gradually outweighed by evidence accumulated from the data, rather than the reverse. Second, it's not clear why, under the efficient communication account, the effect of animacy would be inverted on day 2 (rather than simply being absent). One obvious explanation of the timecourse effect we see in our data is that it reflects some developing understanding of the statistical properties of the data, a point which we return to momentarily.

Third, we think the absence of an effect of event composition is puzzling under the efficient communication account. We saw no difference in the move from inverted-DOM towards a more DOM-like distribution when events were constructed such that unmarked animate objects were actually ambiguous. Recall that the ambiguity of unmarked animate objects is crucial to the efficient communication account of DOM; in the absence of this ambiguity, we would expect case to never be marked, since it is not required to disambiguate. Why would biases which take into account potential ambiguity of linguistic signals be impervious to their actual ambiguity? One possible explanation is participants realise that while an unmarked animate object is not ambiguous *for the set of events they encountered*, it might be ambiguous if the set of potential events to be described were different. This possibility is quite hard to rule out experimentally, but we find it implausible because it mismatches with other experiments we have conducted showing rapid adaptation by participants to the potential ambiguities within the experimental task they are undertaking: for instance, Winters et al. (2015, 2018) show (also using an artificial language paradigm) that participants can rapidly learn that they do or do not need to encode shape or colour of objects in order to

be understood by a communicative partner, and adjust their language use to avoid encoding distinctions which are not relevant to the communicative task they are involved in.

Fourth, and finally, we see a modest increase in DOM in actual communicative interaction: our participants do change their use of case markers when performing a communicative task, casemarking more than they did during non-communicative sentence recall and specifically targeting situations where the absence of ambiguity is problematic, i.e. when the object is animate. This matches the efficient communication account and shows that participants are indeed sensitive to the ambiguity arising from unmarked animate objects: however, this bias appears in communication, where ambiguity matters, rather than in learning.

If efficient communication does not account for the behaviour of our participants during learning, are there other possible explanations which might account for the pattern of effects we see here? We can think of two possibilities which appeal to statistical properties of the language that participants are trained on, without invoking efficient communication.

One possibility is that these results are due to some other learning-driven bias. For instance, it may be that participants are aware that some objects are more typical than others — specifically, inanimate objects are more typical than animate objects — and initially align this with a simple statistical feature of their input, namely that casemarked object nouns are more frequent than unmarked object nouns. If participants initially assumed that the more typical object type (inanimate) should be marked with the more frequent/typical object noun form (featuring the case marker), this would yield an inverted-DOM effect; if participants were subsequently able to learn, as is the case in their input, that object typicality does not align with noun form typicality then this would predict a reduction in this anti-DOM tendency, which is broadly what we see. We will refer to this as the *typicality matching* account, mirroring the term "markedness matching" introduced by Haspelmath (2008). This would constitute a kind of iconicity bias (typicality in semantic space is mirrored by typicality in form space) for which there is independent evidence in artificial language learning (e.g. Culbertson & Adger, 2014). It should be noted that this hypothesis is entirely posthoc, and would therefore require additional supporting evidence. It is also worth noting that typicality matching runs counter to the normal conception of iconicity of markedness matching in linguistics (e.g. Givón, 1991 for the general case, Aissen, 2003 for the same argument applied specifically to differential casemarking), where markedness of a linguistic form is almost always associated with the more weighty form. Here, that would mean presence rather than absence of a case marker (note that markedness matching in our experiment would produce the classic DOM effect, which we do not see in our data). However, when considering markedness in natural languages, frequency and weightiness of form are typically correlated and therefore confounded — weightier forms tend to be less frequent, presumably because of least effort principles. Artificial languages allow us to decouple these two aspects of markedness, as indeed we do here in our input language, where the more weighty casemarked form is more frequent. Again, additional supporting evidence would be required to test the conjecture that it is frequency, rather than weight, which is the relevant factor determining iconicity in our experiment.

A second alternative explanation for our results is that they arise from the interplay of the statistics of the input with other biases in production. As discussed above, our participants were more likely to produce SOV order when the object was inanimate; when both entities in an event were animate, they used OSV order more frequently, consistent with a tendency

to produce animate or human entities before inanimate or non-human entities. Recall that participants saw different frequencies of casemarking for the two word orders: casemarking occurred on two thirds of SOV sentences in their input, but on only half of OSV sentences. If participants were sensitive to this case-order correlation, the combination of this skewing in the input and the observed bias to use SOV order more with inanimate objects would yield an inverted DOM: OSV sentences are more common with animate objects, and casemarked objects should be less frequent in OSV sentences based purely on the statistics of the input, leading to reduced casemarking of animate objects. Since the over-use of OSV order with animates declines over days, we would expect this inverted DOM effect to disappear, which is again broadly the effect we see. This suggests the possibility that at least part of the changing effect of animacy on casemarking could be due to an interplay between the statistics of the input, participants' developing knowledge of those statistics, and an independent bias favouring ordering animates before inanimates.

However, this *case-order correlation* account is not fully supported by our data. Firstly, there is little evidence in their productions that participants learnt the conditioning of casemarking on word order. Figure 8 shows frequency of casemarking by word order, and Table 5 the accompanying statistical analysis; there is no evidence in our participants modulated their frequency of casemarking according to word order on any day. Second, an analysis of casemarking including word order as a predictor (see Table 6) still shows the significant effects of animacy and day x animacy interaction indicating an inverted DOM effect at day 2 that disappears or reverses at day 4 (effect of animacy at Day 2: $b = -0.59$, $SE = 0.22$, $p = .007$; animacy x Day=4 interaction: $b = 0.65$, $SE = 0.19$, $p < .001$)[14], suggesting that the effects of animacy on casemarking are present even when taking into account the effects of word order; note however that this model includes several co-linear predictors (as indicated by multiple Variance Inflation Factors over 3), and there is still no DOM-like effect at day 4 (the effect of animacy is n.s. in a re-levelled model taking Day=4 as the intercept: $b = 0.06$, $SE = 0.21$, $p = .776$). While the case-order correlation account in principle seems viable, there is little evidence to support it in our data.

Experiment 1 therefore provides some evidence that DOM can develop rapidly in communicative interaction (as indicated by a significant shift towards marking animate objects during interaction, seen in our analysis of our participants Day 4 data), but provides much weaker and at best equivocal support for a *learning-driven* efficient communication account of DOM, where the DOM effect manifests itself in non-communicative sentence production tests (as tested by FNJ). Our results are somewhat consistent with existing evidence adduced for that account, in that animacy does influence use of casemarking in non-communicative learning-and-recall tasks. However, there are several features of our data which are puzzling from the efficient-communication-in-learning account, including the temporal profile of the effect (the fact that we see an inverted-DOM configuration on day 2, which unwinds over days) and the fact that it is impervious to actual ambiguity. There are at least two possible alternative explanations for our pattern of results, raised above: these results might be due to

[14]This analysis also shows a marginal animacy by word order interaction ($b = 0.69$, $SE = 0.35$, $p = .048$) which suggests that the effect of animacy at day 2 is most apparent on OSV sentences. It is unclear why this would be the case.

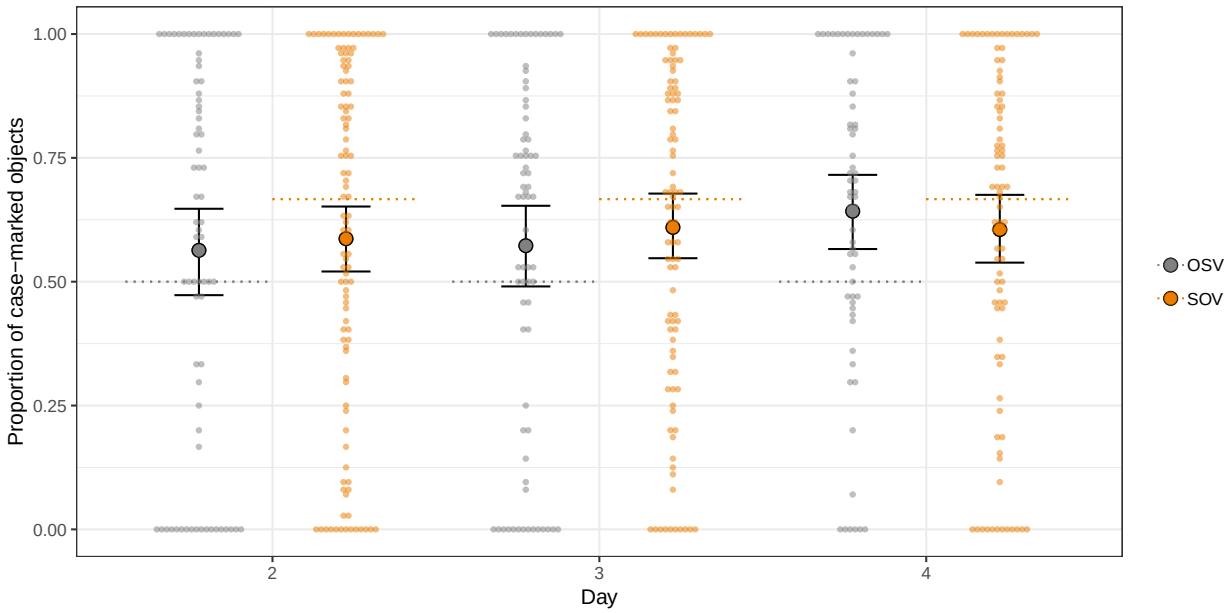

Figure 8: Frequency of casemarking by word order in participants' productions in Experiment 1 (collapsing across event composition). Horizontal dotted lines shows frequency of casemarking in the input. There is little evidence from their productions that participants have learnt that the two word orders are associated with different frequencies of casemarking.

learners simply using the more typical object noun form (casemarked) with the more typical object noun (inanimates), then learning that this alignment is not supported by their data; alternatively, they might be due to statistical features of the input (specifically, that SOV order is more likely to be casemarked) interacting with other production biases (a declining preference for SOV order with inanimate objects). In Experiment 2 we therefore adjust the statistical properties of the input, in ways which should be neutral under efficient communication accounts but which are predicted under either of these alternative explanations to lead to a change in patterns of casemarking.

3. Experiment 2

In Experiment 1 we were concerned that properties of the input (either the fact that casemarked objects are more frequent than unmarked objects, or the details of how casemarking is conditioned on word order) might be responsible for at least some features of our results. In order to investigate this possibility and arbitrate between the efficient communication, typicality matching and case-order correlation accounts, in Experiment 2 we simply invert the relationship between word order and case, changing the language so that OSV order is casemarked more frequently than SOV order in the input; this can straightforwardly be achieved without changing the word-order distribution by flipping from 60% casemarking to 40% casemarking, replacing every occurrence of a casemarked noun in Experiment 1 with an unmarked noun and vice versa. Under this new input configuration, both the typicality matching and case-order correlation accounts now predict a DOM-like effect to be initially

present on Day 2 and decline thereafter.

Under typicality matching, aligning the more typical object type (inanimate) with the more typical noun type (now *unmarked* in the input in Experiment 2) should produce a DOM-like effect on day 2, gradually reducing over days as participants learn this alignment is not present in their input.

Given a tendency to use OSV order with animates (which we observed in Experiment 1) and to casemark OSV sentences *more* frequently (as is the case in the input in Experiment 2), the case-order correlation account would predict participants should immediately casemark animates more, i.e. show an initial DOM effect on day 2; if the tendency to over-use OSV with animate objects declines over days (for which we see support in Experiment 1), we would again expect this DOM effect to reduce over the course of the experiment.

In contrast, under the efficient communication account, changing the frequency of casemarking or the details of the conditioning of case on word order should not substantially affect the emergence of DOM: if efficient communication considerations are at play, we therefore predict the results in Experiment 2 to be broadly similar to those in Experiment 1, i.e. an inverted effect on day 2 (for reasons which remain mysterious), with the language becoming increasingly DOM-like on day 4.

3.1. Method

The method is identical to that used in Experiment 1, the only difference being in the frequency of casemarking in the target language. We again included a between-subject manipulation of event composition (regarding whether subjects could or could not also appear as objects), and again have participants interact communicatively with Smeeble at the end of day 4; in both cases this provides an opportunity to replicate (or not) the modest effect of interaction on DOM we saw in Experiment 1.

3.1.1. Participants

As in Experiment 1, participants were recruited via Amazon Mechanical Turk, and were self-reported native speakers of English aged 18 or over who possessed the MTurk qualification indicating they were based in the US; access to days 2-4 was again controlled using qualifications, with the same criteria as in Experiment 1. The total number of participants participating on each day was: Day 1: 114; Day 2: 90; Day 3: 88; Day 4: 81.[15]

3.1.2. Stimuli and target language

The stimuli and target language were identical to those used in Experiment 1, with the exception of the frequency of casemarking, which was inverted: 40% of objects were casemarked (whereas in Experiment 1, 60% were marked). The grammar of the target language, including the correlations between word order and casemarking arising from inverting the

[15]Of these participants, 2 failed the noun comprehension test on day 1 and do not feature in any of the analyses that follow. 3 failed the sentence comprehension test on day 1 and 1 failed on day 3; we included these participants in the analyses of identification accuracy, but these participants did not complete sentence production trials on the day they failed the sentence comprehension test and did not participate subsequently; therefore they did not contribute data to any of the analyses on casemarking or word order beyond the point at which they failed the sentence comprehension test. Note that the sentence comprehension test is harder when casemarking is less frequent, as there are more ambiguous trials lacking a case marker.

$S \rightarrow N_{subject}\ N_{object}\text{-}ka\ V$ (SOV, marked, $p = 0.2$)

$S \rightarrow N_{subject}\ N_{object}\ V$ (SOV, not marked, $p = 0.4$)

$S \rightarrow N_{object}\text{-}ka\ N_{subject}\ V$ (OSV, marked, $p = 0.2$)

$S \rightarrow N_{object}\ N_{subject}\ V$ (OSV, not marked, $p = 0.2$)

Figure 9: The grammar for Experiment 2. Lexical items and their random assignment are as per Experiment 2, as given in Figure 1.

occurrences of casemarking, are given in Figure 9 — note that as a result of the inversion, non-marking is now more common than marking, and SOV order order is now associated with *less* frequent casemarking than OSV order.

3.1.3. Manipulation of event composition

Identical to Experiment 1.

3.1.4. Procedure

Identical to Experiment 1.

3.1.5. Coding word order and casemarking

Identical to Experiment 1.

3.2. Results

3.2.1. Identification accuracy on comprehension trials

As in Experiment 1, we begin by verifying that our manipulation of the stimuli (whether or not subjects could also appear as objects) had the intended effect of introducing genuine ambiguity for sentences involving two animate participants and no casemarking. The pattern of results here is largely similar to that seen in Experiment 1: the data are presented in Figure 22 in Appendix B, and the statistical analysis is summarised in Table 7. Importantly, performance on trials featuring an unmarked animate object is already high on day 1 in the Subjects Cannot Be Objects condition (significantly higher than 50%, $b = 0.98$, $SE = 0.12$, $p < .001$), and higher than could be achieved by making optimal use of word order cues to agency, $p < .001$), and is significantly higher in subsequent days, indicating again that casemarking of animate objects is increasingly redundant in that condition. By contrast, performance is significantly lower on unmarked animate trials on Day 1 in the genuinely ambiguous Subjects Can Be Objects condition ($b = -0.73$, $SE = 0.16$, $p < .001$), and shows no increase over days.[16]

[16]There are two potentially interesting differences between the results here and the equivalent results for Experiment 1. Firstly, we see significant and positive Casemarking x Day interactions at days 3 and 4; these were n.s. in Experiment 1, presumably due to ceiling effects; in Experiment 2 performance on casemarked nouns is somewhat lower at Day 1, presumably because participants take longer to figure out the function of the less frequent marker, leaving room for a benefit of casemarking at later days. Second, participants

3.2.2. Testing preconditions for the case-order correlation account

One of the three accounts we are evaluating (the case-order correlation account) rests on the assumption that participants will be more likely to use OSV order when the object is animate (as we saw in Experiment 1), and that they are sensitive to the correlation between word order and casemarking in their input (which we had little evidence for in Experiment 1).

Figure 10 shows participants' use of SOV word order over all 4 days, broken down by animacy of the object. We ran a logit regression on participants' word order choices, predicting the probability of SOV order based on day, animacy and their interaction; full details are provided in Table 8. Participants have a substantial preference for SOV word order on day 2, which is significantly higher than the frequency of SOV in the input ($p < .001$); this preference for SOV order was significantly lower at day 4 ($b = -0.82$, $SE = 0.25$, $p = .001$). There was also a significant effect of animacy on word order choice, present at day 2 ($b = -1.28$, $SE = 0.21$, $p < .001$): participants were less like to use SOV order when the object was animate, as we saw in Experiment 1. Further, as in Experiment 1 this tendency to over-use SOV order with inanimate objects declined at days 3 and 4 (Day 3: $b = 0.35$, $SE = 0.17$, $p = .044$; Day 4: $b = 0.47$, $SE = 0.17$, $p = .006$). Figure 11 shows participants' use of casemarking, broken down by word order. As can be seen in the Figure, and as is confirmed by a statistical analysis (Table 9), participants mirror their input in casemarking OSV order more frequently than SOV.

The assumptions underlying the case-order correlation account are therefore met in Experiment 2: we see a preference for OSV order when the object is animate (declining across the 4 days), and participants follow their input in conditioning casemarking on word order. It is however worth noting at this point that the tracking of case-order correlations in Experiment 2 is rather different from what we saw in Experiment 1, where there was little evidence that participants were able to match the conditioning of casemarking on word order present in their input.

3.2.3. Casemarking and animacy during sentence recall

Recall that the three accounts under consideration make rather different predictions about how the change in the statistics of the input language will affect the relationship between casemarking and animacy in participants' productions. The efficient communication account predicts a pattern of results broadly comparable with those seen in Experiment 1; in particular, a growing tendency to preferentially casemark animates (although in Experiment 1 this played out as an initially anti-DOM configuration and culminated in equally frequent marking of animates and inanimates). The typicality matching and case-order correlation accounts predict that the change in the input statistics will result in a change in the effect of animacy. For typicality matching, we expect the opposite pattern from Experiment 1:

show some evidence that they are using word order as a cue to agency (recall that 60% of sentences use SOV order), which we did not see in Experiment 1. In the Subjects Can Be Objects condition, performance is above the level expected under random guessing (50%: $b = 0.25$, $SE = 0.11$, $p = .019$) but not different from what could be achieved by using word order cues (at $p = .160$). This might suggests that the reduction in frequency of casemarking had the effect of making participants more sensitive to word order as a cue to agency, although note that we do *not* see an equivalent effect in Experiment 3.

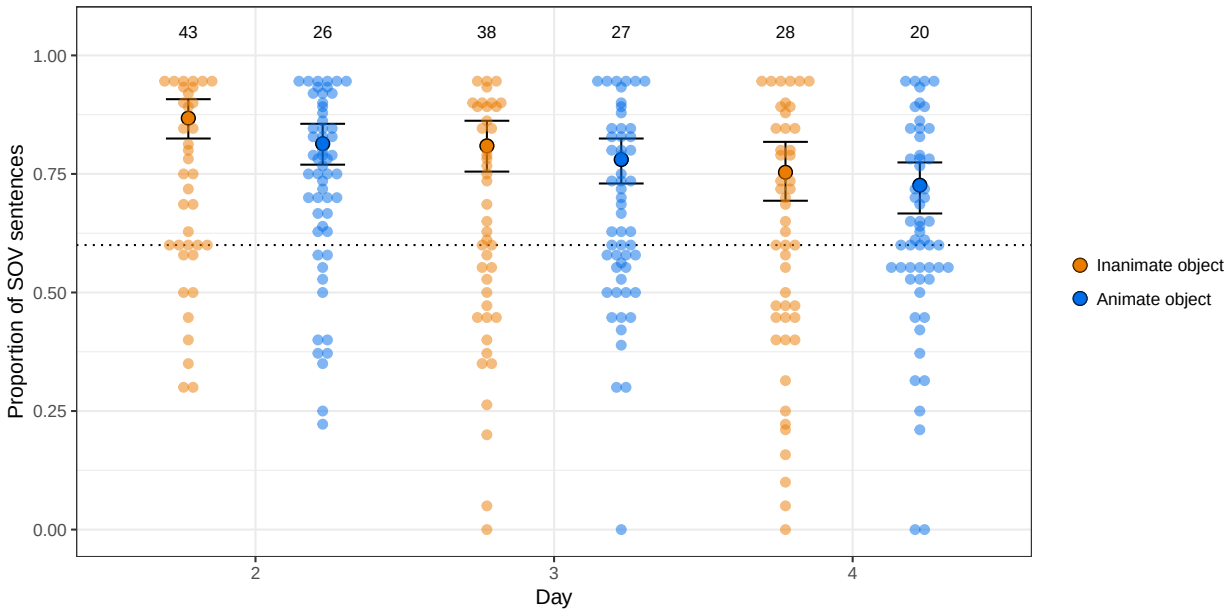

Figure 10: Proportion of SOV (as opposed to OSV) word order in participants' productions in Experiment 2. The horizontal dotted line shows frequency of SOV in the input (for both animate and inanimate nouns). As in Experiment 1, participants over-produce SOV order across all 4 days, and show a preference for SOV order when the object is inanimate (or equivalently, a reduced preference for SOV over OSV when the object is animate).

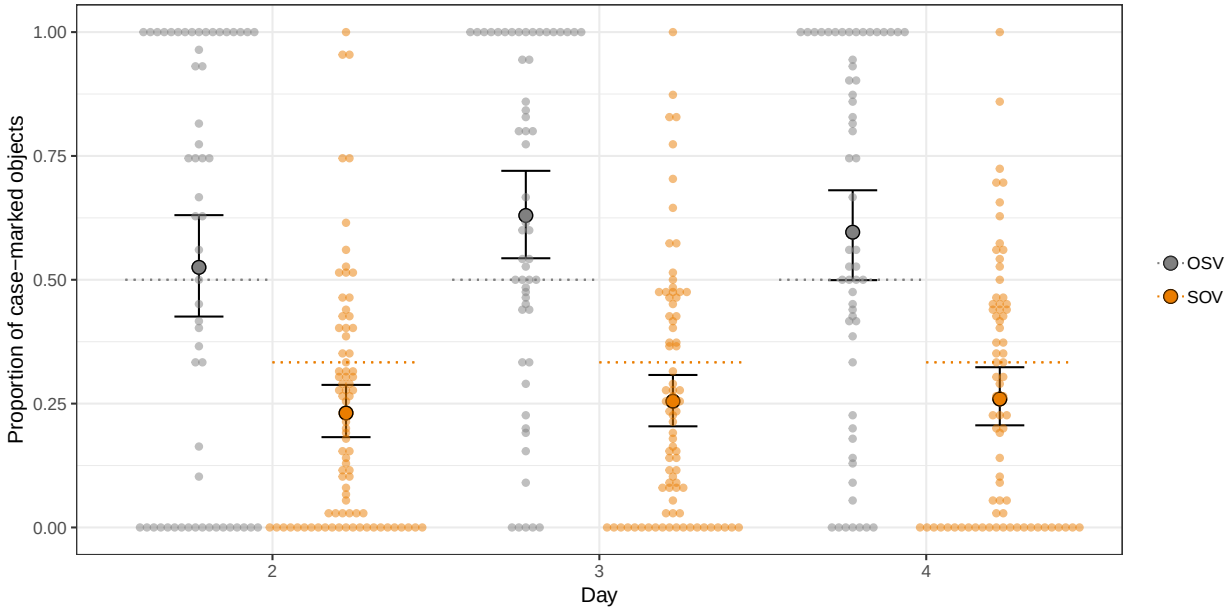

Figure 11: Proportion of casemarking in participants' productions in Experiment 2, broken down by word order. Plotting conventions are as in previous figures; the horizontal dotted line shows case frequency of casemarking in the input, which differs between the two orders. Participants on average reproduce or amplify the conditioning of casemarking on word order present in their input.

more frequent marking of animate objects on day 2 because while animate objects are still atypical, casemarking is now atypical. This effect might decline over days as participants learn from their input not to align in this way. For case-order correlation, based on the tendency to put animates before inanimates and the statistics of the input, we would predict a DOM-like effect starting on day 2 and remaining constant through day 4 (since the analysis above shows that the tendency to use OSV order for animates is relatively constant across days in Experiment 2).

Figure 12 shows the frequency of casemarking for inanimate and animate objects across days 2–4 of the experiment. The logistic regression run on this data (summarised in Table 10) shows a rather different pattern of results from that seen in Experiment 1, and largely an absence of any clear effects. Participants significantly under-produce the case marker at day 2, relative to their training data (estimated log-odds of casemarking = -1.85, significantly lower than the input log-odds of -0.4, $p < .001$). Participants do now mark animate objects more than inanimate objects at Day 2, consistent with DOM, but this effect is only marginal, ($b = 0.38$, $SE = 0.22$, $p = .074$); the effect of animacy does not significantly shift over days (as indicated by n.s. interactions with day, although note the estimated effects here are negative anyway, if anything suggesting a reduction in the DOM-like effect; e.g. Animacy x Day=4, $b = -0.14$, $SE = 0.16$, $p = .357$). Importantly the effect of animacy on day 4 is not significant ($b = 0.24$, $SE = 0.21$, $p = .253$), i.e. there is no evidence of a DOM-like use of casemarking after 4 days.[17]

The results of Experiment 2 are therefore not particularly illuminating in isolation. However, recall that the typicality matching and case-order correlation accounts specifically predict that changing the properties of the input from Experiment 1 to Experiment 2 will produce different uses of casemarking, with more casemarking of animates on day 2 in Experiment 2 (i.e. more evidence of a DOM-like configuration initially) and an absence of the shift towards a DOM-like effect seen over days in Experiment 1. We therefore ran a combined analysis of the production data from both experiments, including Experiment as a fixed effect (along with its interactions with the other fixed effects: see Table 12). This analysis suggests that our manipulation of input frequency had consequences for the relationship between animacy and case, beyond what would be expected by chance: as well as the expected effect of Experiment (indicating less frequent casemarking in Experiment 2 than in Experiment 1), this combined analysis indicates a significantly different effect of animacy between Experiments (as indicated by the significant interactions between Experiment and animacy, $b = 0.72$, $SE = 0.30$, $p = .017$, and a marginal interaction between Experiment, animacy and Day=4, $b = -0.67$, $SE = 0.34$, $p = .053$). Changing the frequency of casemarking in the input and adjusting the correlations between word order and casemarking therefore had an effect, which is not expected under the efficient communication hypothesis.[18]

[17]The other marginal effects are a slightly increased tendency to mark case on Day 3 ($b = 0.42$, $SE = 0.21$, $p = .052$), a slightly increased tendency to mark case on Day 4 in the Subjects Can Be Objects condition ($b = 0.84$, $SE = 0.47$, $p = .072$). Table 11 provides the equivalent analysis including word order as a predictor, which produces the same pattern of null results.

[18]The equivalent analysis including word order as a predictor produces the same pattern of results, as shown in the online analysis materials.

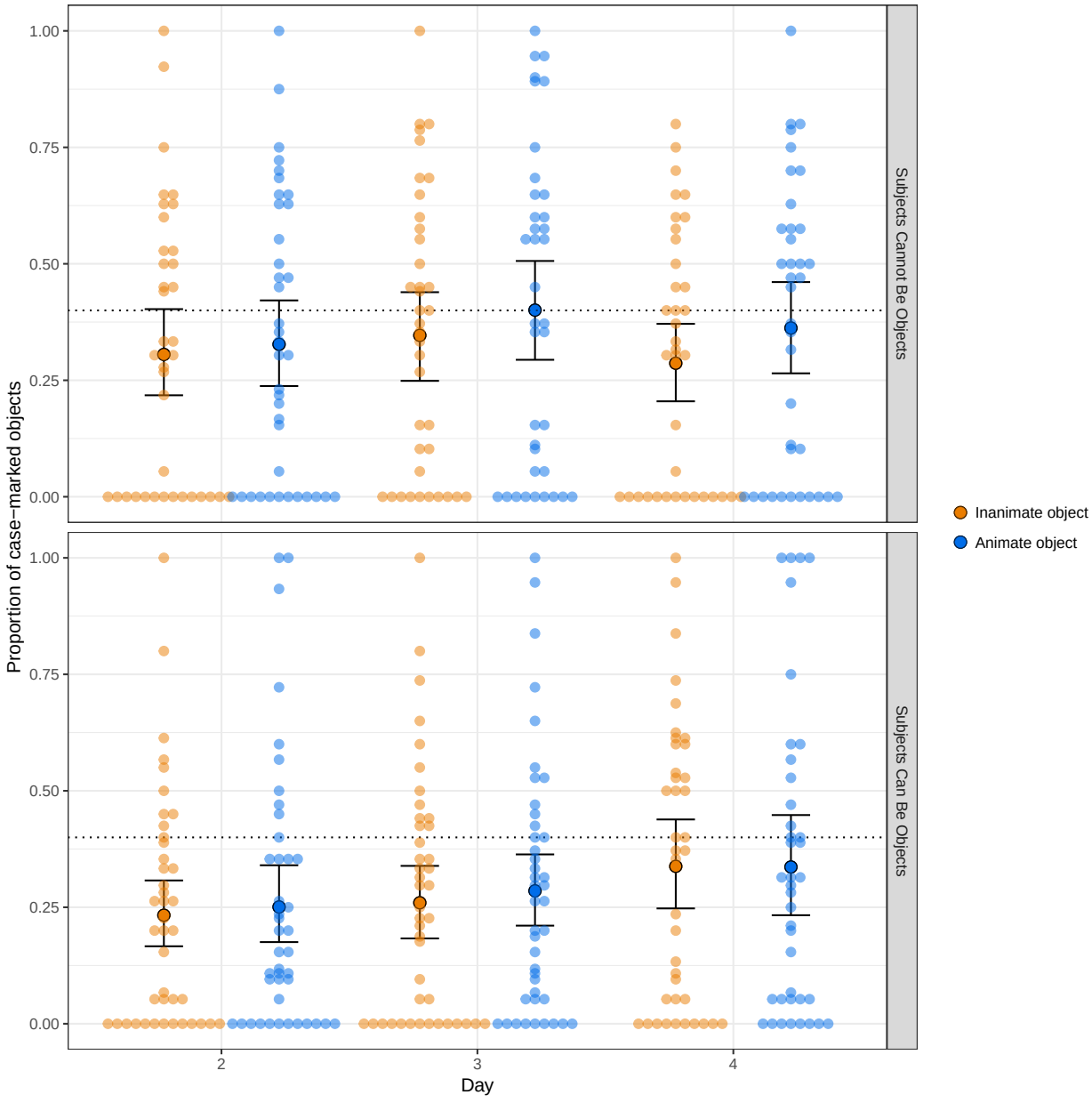

Figure 12: Proportion of casemarking in participants' productions in Experiment 2. The horizontal dotted line shows casemarking frequency in the input (for both animate and inanimate nouns). There is marginally more marking of animates than inanimates at Day 2, and there is no evidence in either event composition that the effect develops over time as we saw in Experiment 1.

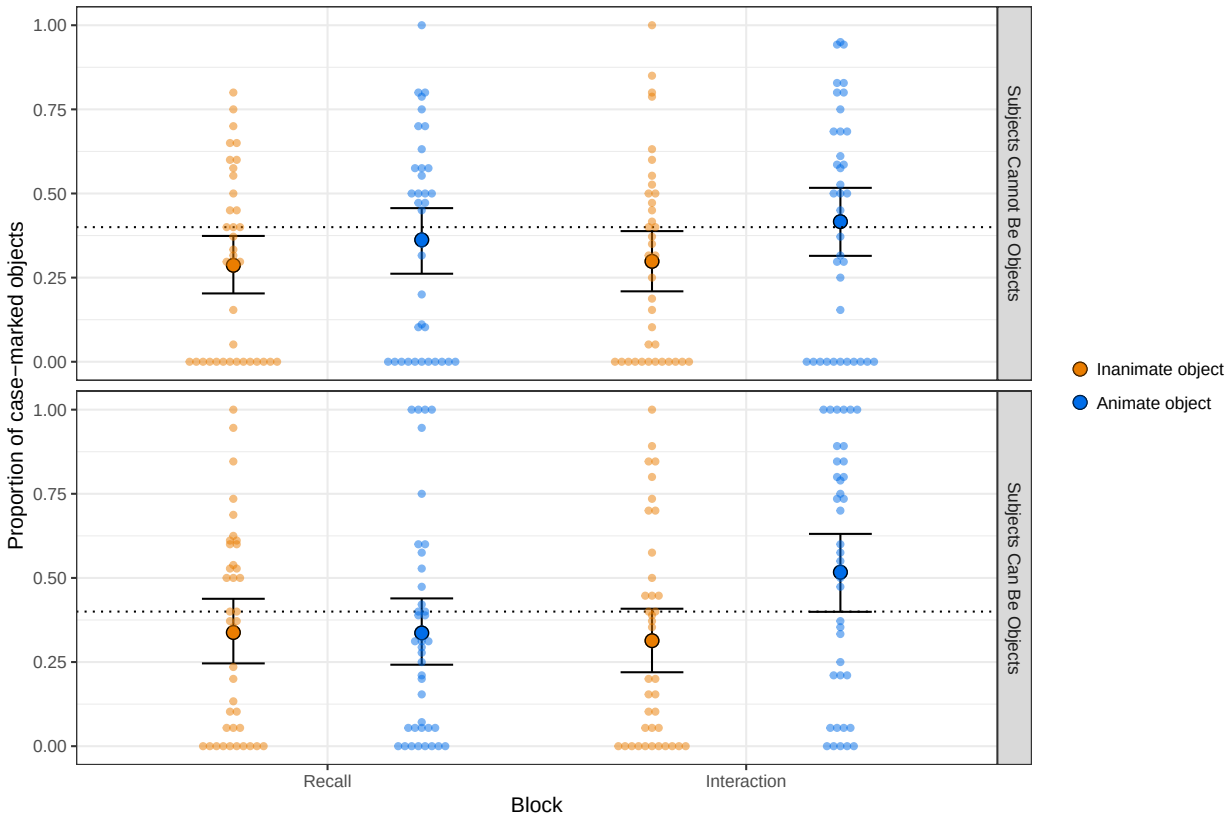

Figure 13: Frequency of casemarking in participants' productions on day 4 of Experiment 2, during the sentence production test (labelled 'Recall' here) and subsequent interaction with Smeeble. The horizontal dotted line shows casemarking frequency in the input (for both animate and inanimate nouns). Participants casemark slightly more during interaction than in the non-communicative recall test, and participants in the Subjects Can Be Objects condition show a DOM-like over-marking of animate objects during interaction that is absent in recall.

3.2.4. Casemarking and animacy during interaction

As in Experiment 1, a visual inspection of communicative accuracy during interaction with Smeeble on day 4 (plot included in online analysis materials) showed that descriptions featuring unmarked animate objects were ambiguous for participants in the Subjects Can Be Objects condition, but were nearly always interpreted correctly by participants in the Subject Cannot be Objects condition; by design, Smeeble found unmarked animate objects ambiguous in both conditions. Figure 13 shows the frequency with which participants used casemarking in their productions during interaction, with their frequency of casemarking on the Day 4 sentence production test also plotted for comparison; Table 13 gives the accompanying statistics.

The interaction phase of Experiment 2 shows a pattern of results broadly in line with our interaction results in Experiment 1. As in Experiment 1, there is no effect of animacy in the pre-interaction recall test on day 4 ($b = 0.30$, $SE = 0.27$, $p = .270$), contrary to accounts where a DOM-like effect should emerge purely through learning (i.e. in non-communicative recall task). As we saw in Experiment 1, there is however a (marginal) increase in the overall

use of casemarking during interaction ($b = 0.36$, $SE = 0.19$, $p = .056$), perhaps reflecting an awareness by participants that failure to casemark might impede communication. Importantly, this effect is greatest for animates (as indicated by significant interaction between animacy and block, $b = 0.98$, $SE = 0.25$, $p < .001$), producing a DOM-like configuration in interaction which was absent at the end of learning. Finally, unlike in Experiment 1, this effect is greatest in the condition where participants might expect that unmarked animates pose the greatest challenge for communication, i.e. the Subjects Can Be Objects condition (as indicated by the interaction animacy, block and event composition; $b = 1.07$, $SE = 0.41$, $p = .009$).

3.3. Discussion

The comparison of Experiment 1 and Experiment 2 shows that changing the properties of the input changes the use of casemarking, as predicted under the typicality matching and case-order correlation accounts, but not under the efficient-communication-in-learning account. Experiment 2 replicates the finding from Experiment 1 that DOM-like effects are absent on day 4 sentence recall, and therefore do not emerge purely through learning as predicted under the efficient-communication-in-learning hypothesis; once again, during learning there is no effect on casemarking of in-practice ambiguity as achieved through our manipulation of event composition, which we think should be predicted under the efficient-communication-in-learning hypothesis. However, it also replicates the Experiment 1 finding that communicative considerations play a role in the use of case markers in actual communicative interaction; both experiments show an increase in casemarking in interaction, which differentially targets the potentially ambiguous case of animate objects, producing a DOM-like configuration during interaction. Experiment 2 therefore serves to cast further doubt on accounts that locate efficient communication effects in non-communicative learning-and-recall tasks; the results do not look as expected under such an account, and are consistent with two alternative accounts that appeal only to participants' developing knowledge of their input statistics, in interaction with other biases acting on learning or production (e.g., a preference for iconicity, or a preference for producing animates first).

However, Experiment 2 does not differentiate between the typicality matching and case-order correlation accounts — both of these predicted roughly the same pattern of results in Experiment 2, either due to the fact that casemarked nouns were atypical (and should be paired with atypical, i.e. animate, objects) or because of case-order correlations in the input (OSV order was casemarked more than SOV). We also noted a mismatch between Experiments 1 and 2 in the ability of participants to reproduce the conditioning of casemarking on word order present in their input: in Experiment 1 participants failed to reproduce the pattern in their input that SOV order was casemarked more frequently than OSV order, whereas in Experiment 2 (where OSV order was marked more than SOV order) participants respected or even exaggerated this aspect of their input. Experiment 3 provides more data on this question, as well as a final pair of input languages where the typicality matching and case-order correlation hypotheses make different predictions.

4. Experiment 3

In order to tease apart the typicality matching and case-order correlation hypotheses, in Experiment 3 we change the structure of the input such that these accounts make different predictions. We do this by again using the languages from Experiments 1–2 as templates, flipping OSV and SOV word orders to look at languages where OSV (not SOV) is the majority order while manipulating frequency of casemarking (40% vs 60% marked objects). This provides us with two additional languages, both of which have 60% OSV order, 40% SOV order, and which differ in the frequency of casemarking. Under the efficient-communication-in-learning hypothesis we would still expect to see either an initial bias towards a DOM-like configuration or the emergence of such a configuration at day 4 for both these languages; however, the typicality matching and case-order correlation hypotheses make different predictions depending on the frequency of casemarking in the input.

In the *40% casemarking* language, casemarking is less frequent than non-marking. Typicality matching therefore predicts an initial DOM-like preference to mark animates (animates are atypical objects, casemarking is atypical). However, because SOV sentences are marked more than OSV sentences, the case-order correlation account predicts an initial anti-DOM configuration (OSV is marked less but should be more likely to be used with animate objects).

In the *60% casemarking* language, casemarking is more frequent than non-marking. Typicality matching therefore predicts an initial anti-DOM configuration (inanimates are typical objects, casemarking is typical). However, because OSV sentences are marked more than SOV sentences, the case-order correlation account predicts an initial DOM-like arrangement (OSV is marked more and should be used more with animate objects).

In both cases we would expect this effect to diminish over days as participants' initial biases are overcome by their training data.

4.1. Method

The method is identical to that used in Experiments 1 and 2, the only difference being in the statistics of the input languages. We again included a between-subject manipulation of event composition (regarding whether subjects could or could not also appear as objects), and again have participants interact communicatively with Smeeble at the end of day 4.

4.1.1. Participants

As in Experiments 1 and 2, participants were recruited via Amazon Mechanical Turk, were self-reported native speakers of English based in the US; access to days 2-4 was again controlled using qualifications, with the same criteria as in Experiment 1. The total number of participants participating on each day was: Day 1: 263; Day 2: 192; Day 3: 175; Day 4: 160.[19]

4.1.2. Stimuli and target language

The stimuli and target language were identical to those used in Experiments 1–2, with the exception of the frequency of SOV/OSV order, which were inverted: 40% of sentences

[19]Of these participants, 9 failed the noun comprehension test on day 1, and none on subsequent days; 21, 2, and 1 failed the sentence comprehension test on days 1–3 respectively, with none failing on day 4.

40% Casemarking language

$S \rightarrow N_{subject}\ N_{object}\text{-}ka\ V$	(SOV, marked, $p = 0.2$)
$S \rightarrow N_{subject}\ N_{object}\ V$	(SOV, not marked, $p = 0.2$)
$S \rightarrow N_{object}\text{-}ka\ N_{subject}\ V$	(OSV, marked, $p = 0.2$)
$S \rightarrow N_{object}\ N_{subject}\ V$	(OSV, not marked, $p = 0.4$)

60% Casemarking language

$S \rightarrow N_{subject}\ N_{object}\text{-}ka\ V$	(SOV, marked, $p = 0.2$)
$S \rightarrow N_{subject}\ N_{object}\ V$	(SOV, not marked, $p = 0.2$)
$S \rightarrow N_{object}\text{-}ka\ N_{subject}\ V$	(OSV, marked, $p = 0.4$)
$S \rightarrow N_{object}\ N_{subject}\ V$	(OSV, not marked, $p = 0.2$)

Figure 14: The grammars for Experiment 3. Lexical items and their random assignment are as per Experiments 1–2, as given in Figure 1.

were SOV (rather than 60% as in Experiments 1–2), and either 40% of objects (as in Experiment 2) or 60% (as in Experiment 1) casemarked. This was achieved by taking the target language from Experiments 1–2 and inverting word order (replacing SOV with OSV and vice versa). The grammar of the target languages is given in Figure 14 — note that in the 40% casemarking language SOV order is associated with more frequent casemarking, and in the 60% casemarking language OSV order is associated with more frequent marking.

4.1.3. Manipulation of event composition

Identical to Experiments 1–2.

4.1.4. Procedure

Identical to Experiments 1–2.

4.1.5. Coding word order and casemarking

Identical to Experiments 1–2.

4.2. Results

4.2.1. Identification accuracy on comprehension trials

As in Experiments 1–2, our manipulation of the stimuli (whether or not subjects could also appear as objects) had the intended effect of introducing genuine ambiguity for sentences involving two animate participants and no casemarking only in the Subjects Can Be Objects condition (see Figure 23 and Table 14): performance on trials featuring an unmarked animate object is already high on day 1 in the Subjects Cannot Be Objects condition ($b = 0.64, SE = 0.06$, significantly higher than 50%, $p < .001$ and indeed higher than could be achieved by making optimal use of word order cues, $p < .001$), and increases significantly in subsequent

days, indicating again that casemarking of animate objects is increasingly redundant in that condition, while in the Subjects Can Be Objects condition performance is at chance, 50% (as indicated by an n.s. intercept in a model with Subjects Can Be Objects as the baseline, $b = -0.03$, $SE = 0.06$, $p = .543$).[20]

4.2.2. Effects of animacy on word order, and tracking of order-marking correlations

Again, since the case-order correlation hypothesis relies on participants over-using OSV order for animate objects and tracking the conditioning of casemarking on order present in their input, we begin by inspecting our data to see if these conditions are met. Figure 15 shows the participants' use of SOV word order over all 4 days, broken down by animacy of the object. We ran a logit regression on participants' word order choices, predicting the probability of SOV order based on Day, Animacy and their interaction (see Table 15 for full details). While there is again a (small) numerical preference to use SOV order more often with inanimate objects than with inanimate objects (i.e. to use OSV order more often with animate objects), this effect is not significant in Experiment 3 ($b = -0.26$, $SE = 0.17$, $p = .121$). This difference from previous experiments may simply be due to relative scarcity of SOV utterances in participants' productions, arising from its low frequency in the input (unlike in Experiments 1–2, SOV order occurred 40% of the time in the input); alternatively, changing the frequency of the word orders in the input may have reduced the preference to place animates before inanimates. This is potentially problematic for evaluating the case-order correlation account, since this relies on the increased tendency to use OSV order with animate objects that we saw in Experiments 1–2.

Figure 16 shows participants' use of casemarking over all 4 days, broken down by word order; Table 16 gives the accompanying analysis. As can be seen in the Figure, and as is confirmed by the statistics, participants appear not to be able to reproduce the conditioning of casemarking on word order present in their input in the 40% casemarking input language (where SOV order was casemarked more than OSV order); there is no effect of word order on casemarking on any day (no effect of word order at day 2, $b = -0.10$, $SE = 0.28$, p .721; no interaction between word order and days 3 and 4, $p \geq .510$); this failure to reproduce the conditioning of casemarking on order is similar to what was seen in Experiment 1, where SOV order was marked more than OSV order. However, participants trained on the 60% casemarking language, where OSV order was casemarked more than SOV order, *were* able to reproduce that conditioning, even on day 2, as indicated by a significant interaction between word order and input language ($b = -1.49$, $SE = 0.39$, $p < .001$); this matches the result of Experiment 2, where OSV order was also casemarked more than SOV order and participants

[20]There is however some suggestion that, as seen in the contrast between Experiments 1–2, participants trained on the language with less frequent casemarking may be more attuned to word order as a cue to agency: this is seen in lower performance on unmarked items on day 1 in the 60% casemarking language ($b = -0.23$, $SE = 0.11$, $p = .039$; the interaction with Subjects Can Be Objects is n.s., suggesting a similar effect in both conditions); the magnitude of this difference also seems to grow with continued exposure to the language (as indicated by a significant interaction between casemarking frequency and day=4: $b = -0.91$, $SE = 0.36$, $p = .012$; however, even at day 4, participants in the Subjects Can Be Objects condition are not performing above chance, suggesting this benefit is small and variable (as indicated by an intercept n.s. different from 0 in a re-levelled model taking Subjects Can Be Objects, Day=4 as the intercept: $b = .08$, $SE = .24$, $p = .720$).

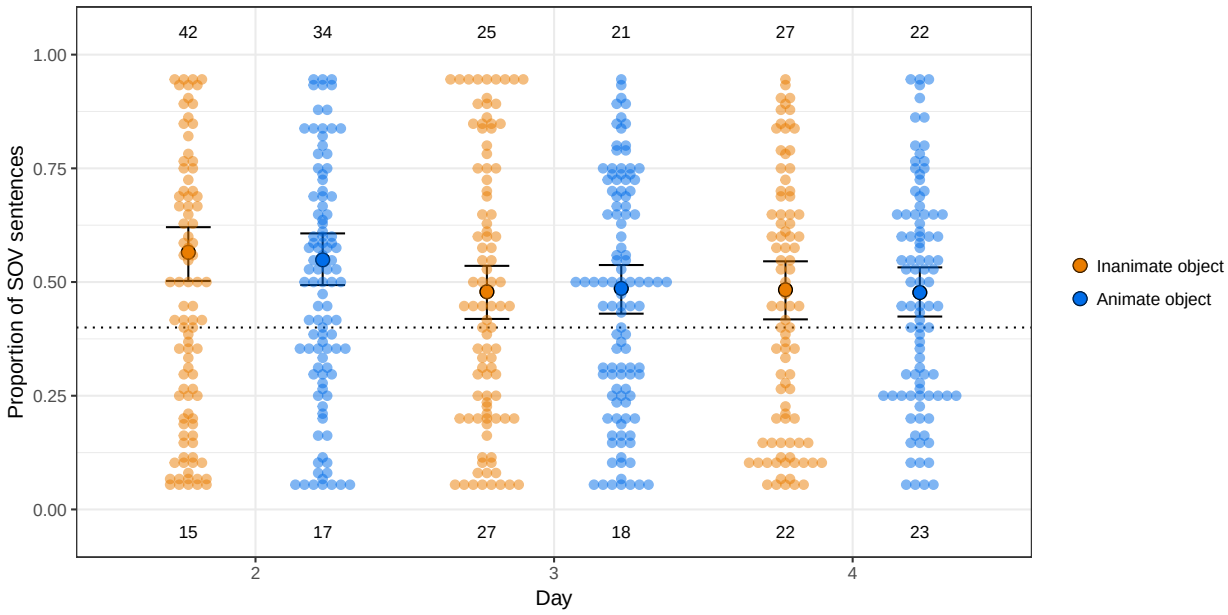

Figure 15: Proportion of SOV (as opposed to OSV) word order in participants' productions in Experiment 3. Text annotations above/below show the number of participants producing all-SOV/all-OSV data since they are too numerous to plot individually. The horizontal dotted line shows case frequency of SOV in the input (for both animate and inanimate nouns). As in Experiment 1–2, participants over-produce SOV order across all 4 days, although this preference declines with more training; there is a slight numerical preference for SOV order when the object is inanimate (or equivalently, a reduced preference for SOV over OSV when the object is animate) on day 2, but this is less pronounced than in Experiments 1–2 and not statistically significant.

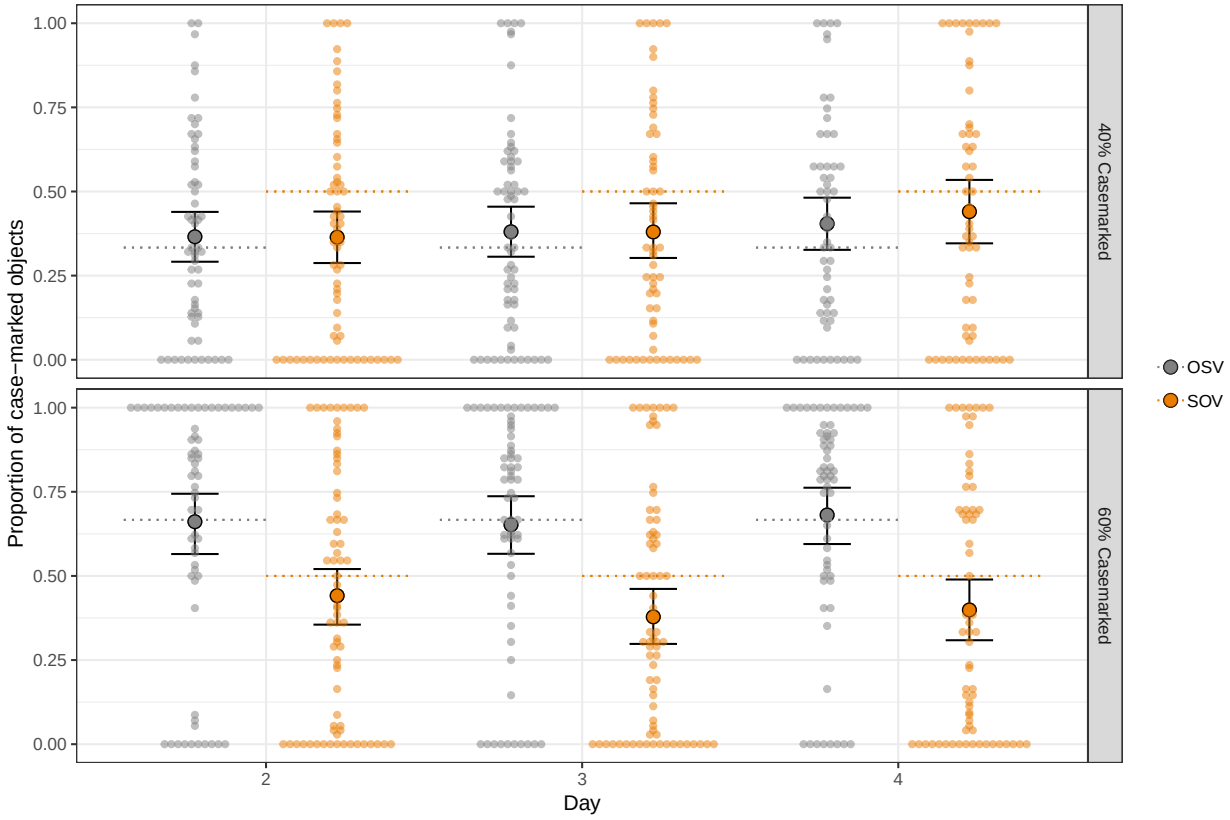

Figure 16: Proportion of casemarking in participants' productions in Experiment 3, broken down by word order. The horizontal dotted line shows case frequency of casemarking in the input, which differs between the two orders. In the 40% casemarking language participants fail to reproduce the conditioning of case on word order present int heir input (mirroring Experiment 1); in the 60% casemarking language participants successfully reproduce or amplify this conditioning (mirroring Experiment 2).

were able to reproduce this in their output. We return to possible reasons for this selective failure to reproduce the conditioning of casemarking on word order in the discussion; for now we simply note that the assumptions of the case-order correlation account are not met in Experiment 3.

4.2.3. Casemarking and animacy during sentence recall

Figure 17 shows the frequency of casemarking for inanimate and animate objects across days 2–4 of the experiment, broken down by our manipulation of event composition and input language. The results do not seem to clearly fit the predictions under *any* of the accounts under consideration. Contrary to the predictions of the efficient-communication-in-learning hypothesis, we do not see any consistent DOM-like effects of animacy at day 2 or day 4. The results in some cells are as predicted by the typicality matching account (e.g. an anti-DOM pattern on day 2 in the 60% casemarked language, but only in the Subjects Can Be Objects condition, a more DOM-like arrangement in day 2 in the 40% casemarked language, but only in the Subjects Cannot Be Objects condition) and show a trend in the opposite direction elsewhere (although notably flatter). While the case-order correlation account is already

dispreferred (not least because of the strong indication that some other factor intervenes in participants ability to reproduce the conditioning of case on word order which it assumes), it shows the same mix of consistent and inconsistent results, with the clearest pattern in our data (the anti-DOM configuration in the 60% casemarking language, at least for Subjects Can Be Objects events) being in the opposite direction to the prediction. Indeed, the figure suggests that the initial pattern of casemarking may be more strongly determined by the event composition, with the event composition which should lead to the greatest preference for a DOM-like configuration (Subjects Can Be Objects) showing the clearest *anti*-DOM arrangement.

The logistic regression run on this data (summarised in Table 17) shows the expected clear effect of the proportion of casemarked objects in the input but no consistent effect of animacy at day 2 ($b = -0.07$, $SE = 0.17$, $p = .669$) and no consistent change in the effect of animacy at day 4 ($b = -0.09$, $SE = 0.18$, $p = .619$), i.e. no evidence for a bias in learning favouring DOM, either on day 2 or developing over days. The interaction between the input language (i.e. the proportion of casemarked objects in the input) and animacy is in the direction predicted by the typicality matching account, but n.s. ($b = -0.54$, $SE = 0.33$, $p = .102$). Instead, there is an unexpected interaction between event composition and animacy ($b = -0.67$, $SE = 0.32$, $p = .036$), indicating that animate objects are *less likely* to be casemarked than inanimate objects at day 2 in the Subjects Can Be Objects condition, i.e. an anti-DOM effect where the effects of efficient communication might reasonably be expected to be at their greatest. There is also a significant interaction between event composition and proportion of casemarked objects ($b = 2.48$, $SE = 0.88$, $p = .005$), reflecting the fact that there appears to be a larger effect of event composition in the 60% casemarking condition than the 40% casemarking condition.[21] Model comparison confirms the equivocal nature of the data: models featuring animacy and proportion (representing the efficient communication account) and animacy, proportion and their interaction (representing the typicality matching account) do not significantly differ in their fit to the data ($\chi^2(3) = .270$).

4.2.4. Casemarking and animacy during interaction

As in Experiments 1–2, an analysis of communicative success during the interaction phase on day 4 revealed that descriptions featuring unmarked animate objects were only ambiguous for our participants in the Subjects Can Be Objects condition, and were nearly always interpreted correctly by participants in the Subject Cannot be Objects condition, while Smeeble (by design) found unmarked animate objects ambiguous in both conditions. Figure 18 shows the frequency with which participants used casemarking on their productions during interaction, with their frequency of casemarking on the Day 4 sentence production test also plotted for comparison; Table 18 gives the accompanying statistics.

As in Experiments 1–2, there is no effect of animacy in the pre-interaction recall test on day 4 ($b = -0.04$, $SE = 0.16$, $p = .808$), but there is an overall increase in the use of casemarking during interaction ($b = 0.62$, $SE = 0.16$, $p < .001$). This effect is again greatest for animates (as indicated by significant interaction between Animacy and Block, $b = 0.61$, $SE = 0.17$, $p < .001$), producing a DOM-like configuration in interaction which was absent

[21] A more complex model including word order as a predictor produces broadly the same pattern of results, see the supporting online analyses.

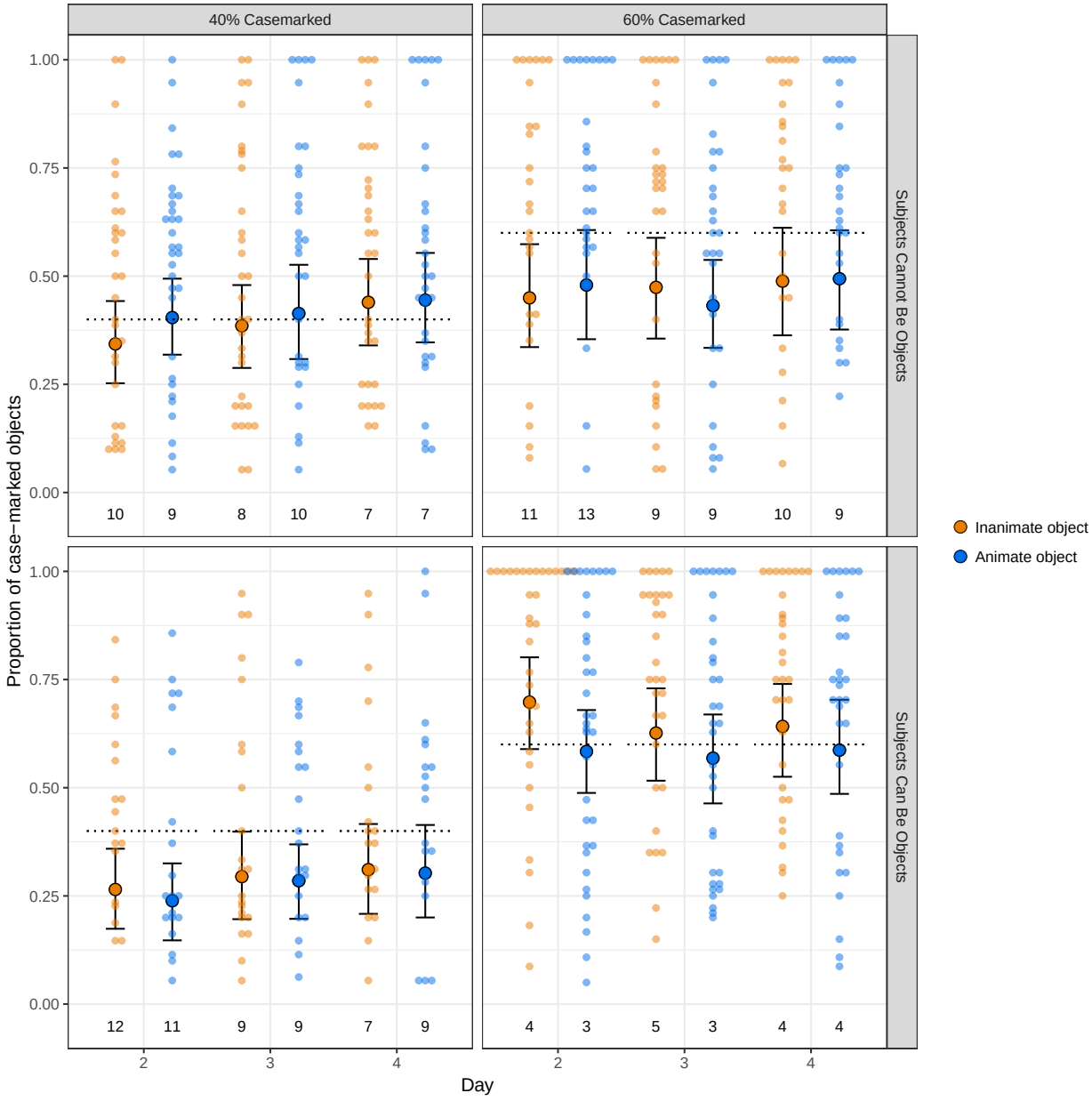

Figure 17: Proportion of casemarking in participants' productions in Experiment 3. The horizontal dotted lines shows casemarking frequency in the input (for both animate and inanimate nouns). The results do not clearly follow any of the predicted patterns, with use of case markers apparently modulated by event composition and animacy rather than animacy alone or animacy in interaction with the proportion of casemarking in the input language.

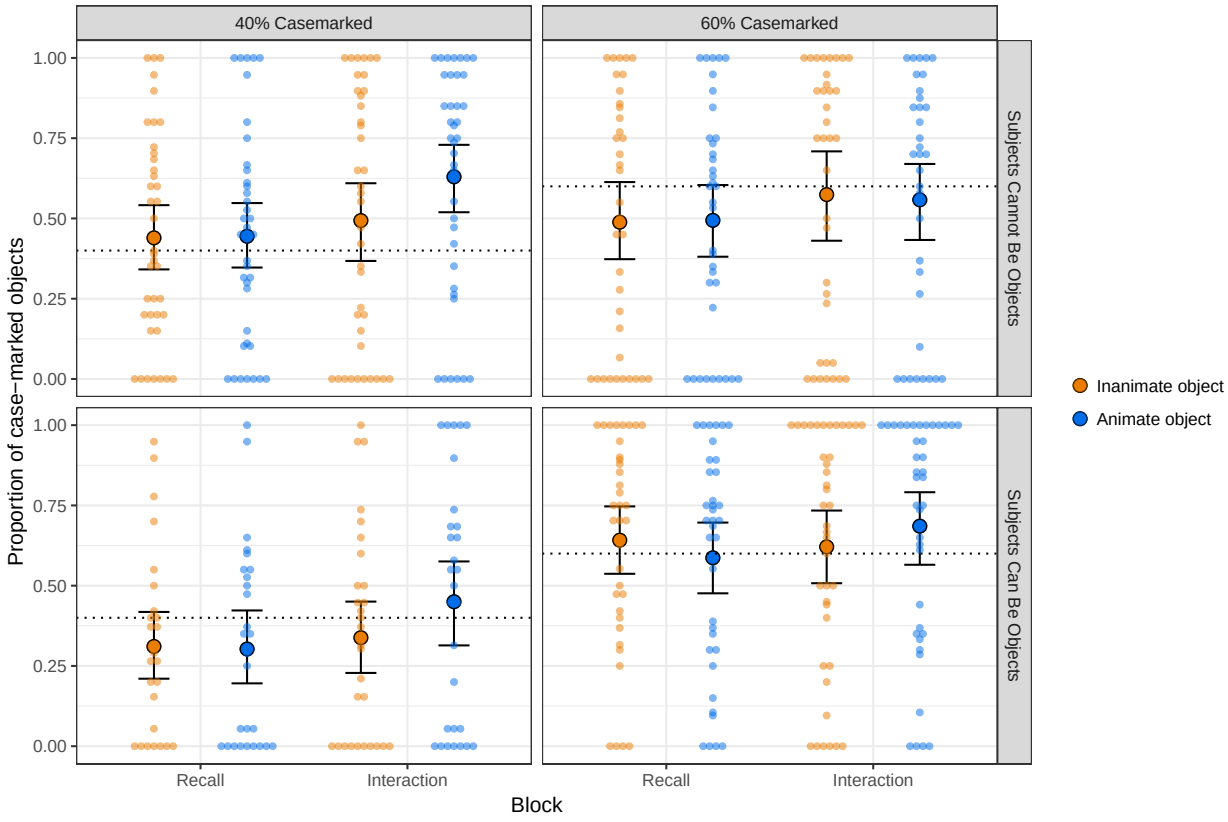

Figure 18: Frequency of casemarking in participants' productions on day 4 of Experiment 3, during the sentence production test (labelled 'Recall' here) and subsequent interaction with Smeeble. The horizontal dotted line shows casemarking frequency in the input (for both animate and inanimate nouns). Participants casemark more during interaction than in the non-communicative recall test, and in most cells show a DOM-like over-marking of animate objects during interaction that is absent in recall.

at the end of learning; like Experiment 1 but unlike Experiment 2, this effect is of the same magnitude in the Subjects Cannot Be Objects and Subjects Can Be Objects conditions (as indicated by an n.s. interaction between animacy, block and event composition; $b = 0.39$, $SE = 0.32$, $p = .228$). Once again, we see no evidence of a DOM-like preference to over-mark animate objects after 4 days of training during the non-communicative recall test, but this configuration rapidly emerges in actual communicative interaction.

4.3. Discussion

Experiment 3 once again replicates our finding in Experiments 1–2 that, after 4 days of training, participants do not produce a DOM-like configuration during non-communicative sentence recall. There is no consistent evidence for a DOM-like over-marking of animate objects anywhere, and in the conditions where we expect to see these effects at their strongest (e.g. at day 2 in the Subjects Can Be Objects conditions) we see the best evidence for *anti*-DOM effects, where participants under-mark animate objects. In short, we see very little support for the efficient-communication-in-learning hypothesis in our data. By contrast, participants rapidly switch to preferentially marking animate objects during actual communicative interaction, where ambiguity matters.

We also sought to further test two additional hypotheses suggested by results from Experiments 1 and 2. However, participants' behaviour across 4 days of learning in Experiment 3 was not as predicted under either of these theories. Recall that under the case-order correlation account, participants' behaviour should be influenced by a preference to produce animates first (i.e. to use more OSV with animates), combined with sensitivity to the input in terms of the relationship between casemarking and word order. We believe that this hypothesis can be ruled out. Firstly, participants' sensitivity to the relationship between case and word order in their input seems to be modulated by some other factor. We consistently find that participants are unable to reproduce the case-order correlation present in their input when it required them to mark SOV order more than OSV order (Experiment 1; Experiment 3 in the 40% casemarking language); however, when the case-order correlation involves marking OSV order more than SOV order, it is successfully reproduced (Experiment 2; Experiment 3 in the 60% casemarking language). The central assumption of the case-order correlation account is therefore not met. We return in the General Discussion to why participants might show this selective failure to reproduce case-order correlations present in their input. Furthermore, the only significant effect of animacy on casemarking we saw in Experiment 3 was in the opposite direction to the prediction of the case-order correlation hypothesis: participants in the 60% casemarking language in the Subjects Can Be Objects condition actually showed a tendency to produce more casemarking with *inanimates* even though under the case-order correlation hypothesis their input language should have encouraged the opposite (more casemarking on OSV, and therefore more casemarking with animates).

The typicality matching account also does not do a good job of predicting participants' use of casemarking in this experiment. There is some evidence consistent with this account, in that the clearest effects of animacy on casemarking seen in Figure 17 are in the direction predicted by typicality matching. In particular, we see DOM-like preferential marking of animates on day 2 in the 40% casemarking condition, where animate objects and casemarking

are both atypical (seen in the Subjects Cannot Be Objects condition) but an anti-DOM-like preferential marking of inanimates on day 2 in the 60% casemarking condition, where inanimate objects and casemarking are both typical (seen strongly in the Subjects Can Be Objects condition). However, these effects are not consistent and are reversed elsewhere. Instead, event composition (whether or not subjects ever appear as objects) seems to have some impact on participants' tendency to mark animate objects. We return to this puzzling result in the General Discussion.

The results of Experiment 3 therefore do not allow us to pinpoint the mechanism driving participants' behaviour in the learning portion of the task: they provide further evidence against FNJ's efficient-communication-in-learning hypothesis, are inconsistent with one of our alternatives (case-order correlation), and fail to provide good support for the remaining alternative, typicality matching.

5. Combined analysis of Experiments 1–3

Experiments 1–3 provide a substantial sample of data (N=341 participants on day 4), an order of magnitude larger than that used in most artificial language experiments, including FNJ. While our participants are clearly highly variable, this large sample should in principle give us a chance to spot relatively small effects which each experiment taken individually might miss. We therefore conduct an analysis on the combined dataset, focusing on 3 questions related to efficient communication and typicality matching: 1) Do we see DOM-like effects at day 2 or day 4, as predicted under efficient communication, or does animacy interact with the proportion of casemarked objects in the input, as predicted under typicality matching?[22] 2) Do we see DOM-like effects strengthen or appear in actual communicative interaction with Smeeble? 3) Is participant behaviour in interaction modulated by our manipulation of event composition?

5.1. Typicality matching beats efficient-communication-in-learning

The combined data from Experiments 1–3 provide a further opportunity to test whether the efficient-communication-in-learning or typicality matching theories better account for our participants' behaviour: as well as manipulating animacy of the object (within-subjects) and casemarking frequency (between-subjects), across the three experiments we also have the full factorial combination of which word order is casemarked more frequently in the input (SOV or OSV) and event composition (Subject Cannot/Can Be Objects). Figure 19 shows the effect of animacy on casemarking, collapsing across all 3 experiments and split by input proportion of casemarked objects (as this is predicted to modulate the effects of animacy under the typicality matching account); Table 19 gives the summary table for the corresponding statistical analysis (featuring all four between-subjects predictors).

[22] We also looked in the combined dataset at whether participants' word order was influenced by animacy. An analysis of the combined dataset, taking use of SOV order (as opposed to OSV order) as the binary dependent variable, reveals a significant effect of animacy ($b = -0.72$, $SE = 0.12$, $p < .001$) on order in the expected direction at day 2 (i.e. less use of SOV order with animate objects), and marginal interactions between animacy and day at days 3 ($b = 0.26$, $SE = 0.14$, $p = .065$) and 4 ($b = 0.25$, $SE = 0.14$, $p = .077$) which suggesting this tendency may decline at later days of training.

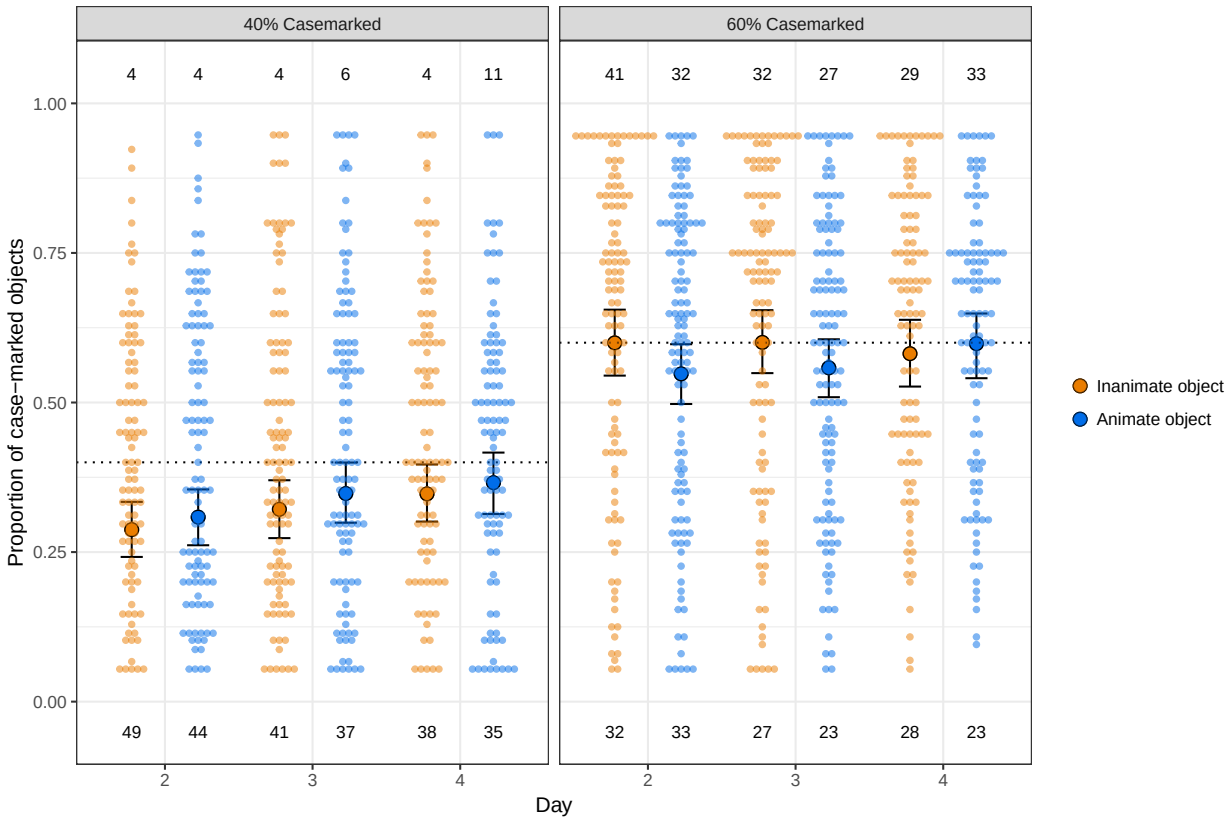

Figure 19: Proportion of casemarking in participants' productions across all 3 experiments. The horizontal dotted lines shows casemarking frequency in the input (for both animate and inanimate nouns). Contrary to the predictions of the efficient-communication-in-learning hypothesis, there is no consistent relationship between animacy and casemarking, and no reliable tendency to over-mark animate objects on day 4; however, as predicted by the typicality alignment hypothesis, there is an interaction between animacy and frequency of casemarking in the input on day 2, with animate objects receiving atypical casemarking (i.e. more marking than animates in 40% Casemarked input languages, less marking than inanimates in 60% Casemarked input languages).

As is clear from the figure and verified in the statistics, across all our data there is no consistent DOM-like bias in production data at day 2 (as indicated by an n.s. effect of animacy, $b = -0.07$, $SE = 0.11$, $p = .523$) and no interaction between animacy and day=4 which would indicate a shift towards a more DOM-like configuration over 4 days ($b = 0.14$, $SE = 0.12$, $p = .258$). Furthermore, there is no interaction between animacy and event composition, which we think would be a reasonable prediction under the efficient communication account (this would predict a positive interaction, i.e., more marking of animate objects in the Subjects Can Be Objects condition; instead we see a negative but n.s. interaction, $b = -0.31$, $SE = 0.22$, $p = .148$). This pattern of results strongly suggests that our data—for the pre-interaction, learning-and-recall portion of the experiment—does not match the predictions of the efficient-communication-in-learning hypothesis.

However, the results taken collectively do match the predictions of the typicality matching account; in particular, there is an interaction between animacy and proportion of casemarked objects ($b = -0.63$, $SE = 0.22$, $p = .005$) which indicates, on day 2, a DOM-like configuration (more marking of animate objects) when casemarking is infrequent in the input, but an anti-DOM configuration (more marking of inanimate objects) when casemarking is frequent, as would be predicted if participants were using typical marking with typical objects and atypical marking with atypical objects. Furthermore, this effect is eliminated by day 4, as participants initial biases are overwhelmed by the evidence in their data that typicality and marking are not aligned in this way (as indicated by a significant three-way interaction between animacy, proportion of casemarking in the input, and day=4: $b = 0.54$, $SE = 0.25$, $p = .029$).[23] Finally, fitting two simpler models representing the two competing hypotheses[24] indicates that the model capturing the core feature of typicality matching, namely that the animacy of the object should interact with the frequency of casemarking in the input, provides a significantly better fit to the data ($\chi^2(3) = 14.93, p = .002$).

5.2. Differential Object Marking develops in interaction

Figure 20 shows day 4 data across all 3 experiments, contrasting the (non-communicative) pre-interaction recall test with behaviour during interaction; Table 20 gives the accompanying stats. As can be seen in Table 20, aggregating over all 314 participants who provide usable day 4 data, we see no reliable effect of animacy on casemarking during sentence recall ($b = 0.15$, $SE = 0.11$, $p = .188$); however, there is a substantial and significant increase in casemarking during interaction ($b = 0.66$, $SE = 0.11$, $p < .001$), and an increased tendency to mark animate objects (as indicated by a significant block x animacy interaction: $b = 0.68$, $SE = 0.22$, $p < .001$); in other words, while there is very little evidence across our 3 experiments that a DOM-like configuration emerges from learning, it rapidly develops during communication when the ambiguity introduced by casemarking actually matters.

This analysis also reveals a small but significant increase in differential casemarking in

[23] We see the same key results in an analysis including participants' word order as an additional predictor; see the supporting online analyses

[24] In lme4 syntax, the efficient-communication-in-learning model is `casemarking ~ day * (animacy + input proportion)` and the typicality matching model is `casemarking ~ day * animacy * input proportion`, i.e. they differ only in whether animacy interacts with the proportion of casemarked objects in the input.

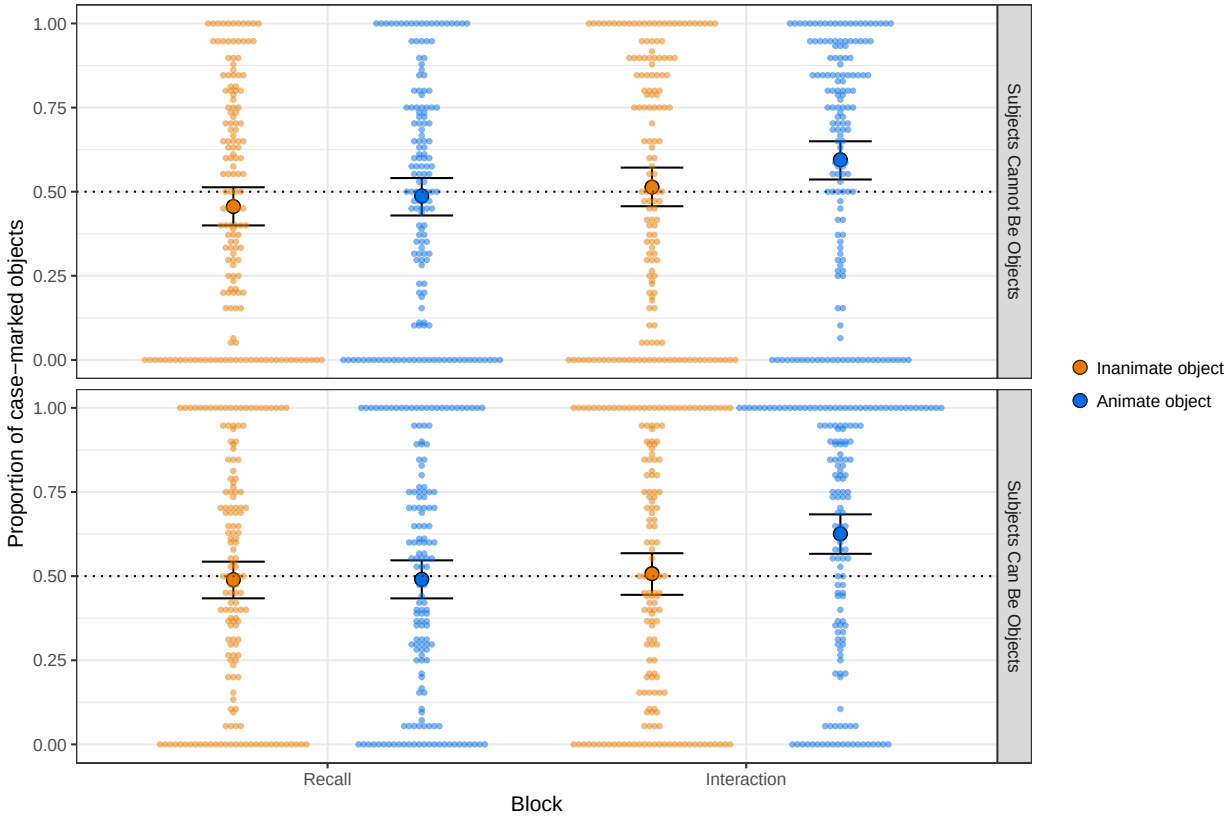

Figure 20: Frequency of casemarking in participants' productions on day 4 across all 3 experiments, during the sentence production test (labelled 'Recall' here) and subsequent interaction with Smeeble. The horizontal dotted line shows casemarking frequency in the input (for both animate and inanimate nouns). While there is no effect of animacy during non-communicative recall, participants show a DOM-like over-marking of animate objects during interaction, and this effect is larger in the Subjects Can Be Objects condition, where unmarked animate objects result in genuine ambiguity for participants.

the Subjects Can Be Objects condition (as indicated by a significant three-way interaction between block, animacy and event composition: $b = 0.52$, $SE = 0.21$, $p = .016$), suggesting that participants increase their use of differential marking in the condition where, in their experience, unmarked animates pose genuine ambiguity. However, this effect is clearly highly variable and only comes out in our combined analysis (and in the analysis of Experiment 2 interaction data). Recall that for our participants, unmarked animates were only ambiguous in the Subjects Can Be Objects condition; by day 4, virtually all participants in the Subjects Cannot Be Objects condition had learned to exploit the structure of the event set when interpreting unmarked animate objects. However, due to the way we modelled Smeeble's behaviour when playing the role of matcher, unmarked animate objects were problematic for Smeeble in both conditions: since we weren't certain prior to conducting Experiment 1 than participants would find unmarked animates unambiguous in the Subjects Cannot Be Objects condition, we didn't want to build this into their interlocutor. The fact that this interaction involving event type is significant suggests that participants' own experience of the ambiguity of unmarked animate objects has an influence, although the presence of an effect of animacy in interaction even in the Subjects Cannot Be Objects condition suggests that participants' behaviour in interaction was largely adapted to the requirements of their interlocutor.

This is confirmed by an analysis of the timecourse of participants' use of casemarking during interaction, shown in Figure 21: while participants in the Subjects Can Be Objects condition mark animate objects more than inanimate objects from the start of interaction, suggesting immediate adaptation to the requirements of the communicative task, participants in the Subjects Cannot Be Objects condition show less of a tendency to do so (because in their experience unmarked animates are not problematic), and adapt to Smeeble's requirement for animate objects to be marked over the course of the interaction. This impression is (weakly) confirmed by the statistical analysis (see Table 21): in the Subjects Cannot Be Objects condition there is no clear initial effect of the object's animacy on casemarking ($b = 0.40$, $SE = 0.23$, $p = .082$ — note that this is roughly the same degree of differentiation that these participants showed at the end of their sentence production test, where the effect of animacy was n.s.[25]), and while the overall level of casemarking does not increase over time (as indicated by an n.s. effect of trial number) there is a tendency for animate and inanimate objects to be increasingly differentiated (as indicated by a marginal positive interaction between animacy and trial number: $b = 0.015$, $SE = 0.008$, $p = .054$). In contrast, in the Subjects Can Be Objects condition, animate and inanimate objects are differentiated from early on (the interaction between event composition and animacy is significant: $b = 0.69$, $SE = 0.33$, $p = .039$), and this differential marking does not increase in the same way it does in the Subjects Cannot Be Objects condition (as indicated by a significantly negative 3-way interaction between animacy, event composition and trial number, $b = -0.02$, $SE = 0.01$,

[25]In the supporting online analyses we provide plots and inferential statistics which verify that the modest numerical difference seen between inanimate and animate objects at the end of the recall test seen in the Subjects Cannot Be Objects condition in Figure 21 is purely a product of chance fluctuation in marking frequency: an analysis looking at the effect of animacy and trial number shows no reliable interaction, i.e. this effect does not develop over time during recall (e.g. due to repeated testing) in the same way as it does during interaction.

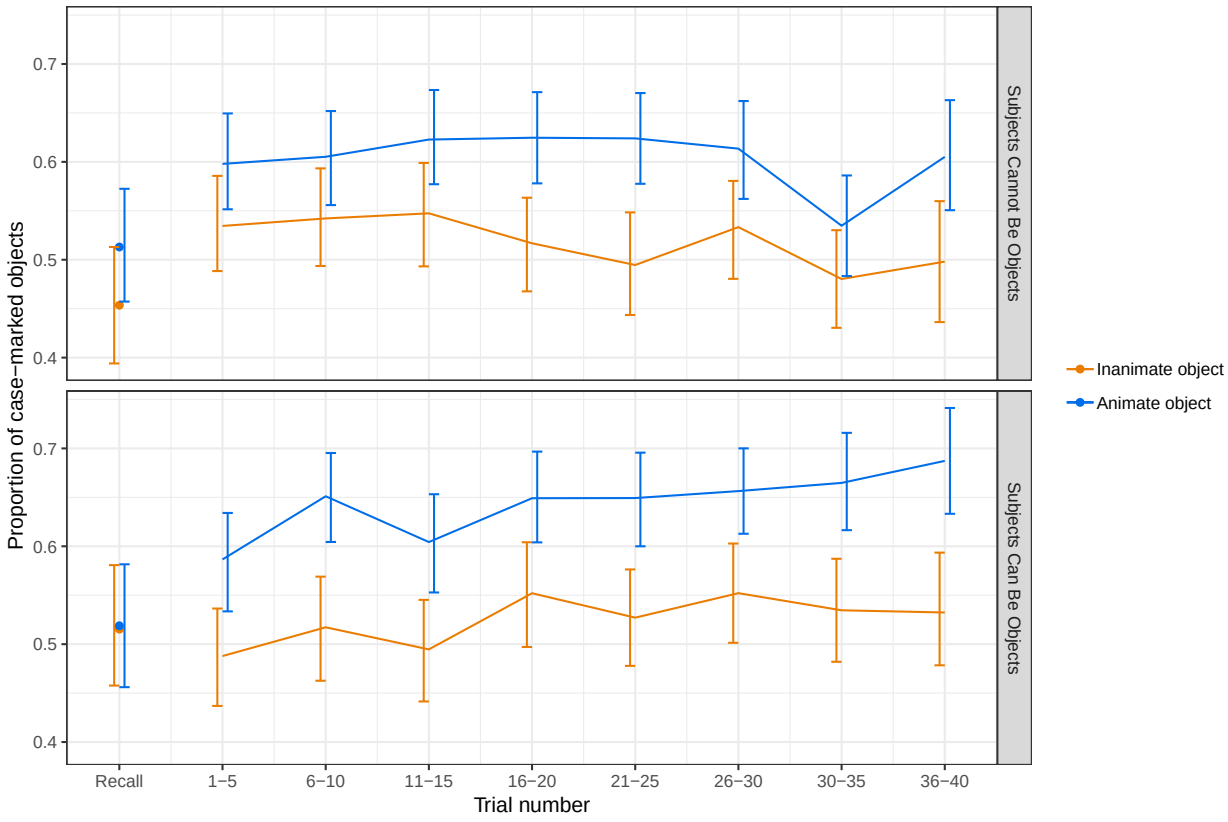

Figure 21: The timecourse of participants' use of casemarking during their interaction with Smeeble, split by event type: points labelled as Recall show means in the last 5 trials of the non-interactive recall test, lines show means over time during interaction, error bars indicate bootstrapped 95% confidence intervals. In the Subjects Cannot Be Objects condition participants do not show a preference for marking animate objects early in interaction (since they themselves do not find unmarked animates ambiguous); however, over the course of their interaction with Smeeble (who *does* find unmarked animates ambiguous) they show a weak tendency to differentiate, marking animate objects more than inanimates. In the Subjects Can Be Objects condition (where participants find unmarked animate objects ambiguous), this DOM-like pattern is present from the start of interaction.

$p = .035$).

5.3. Discussion of combined analyses

In sum, across our 3 experiments we see only fleeting evidence for any effect of animacy on casemarking during learning; this effect is modulated by the frequency of casemarking in the participants' input, and extinguished after 4 days of training. Participants' behaviour during learning is better explained by typicality matching than the efficient-communication-in-learning account proposed by FNJ.[26] We do however see very clear effects of communicative

[26]Note that we see quite clear effects of animacy on word order during learning noted in footnote 22 above. This, in combination with the fact that participants distinguish between animates and inanimates during actual communicative interaction, strongly suggests that the animate/inanimate contrast was a salient feature of our stimuli for our participants, and sufficient to elicit effects on word order if not on casemarking;

use of the language on casemarking: participants rapidly adjust their use of casemarking in communication, quickly switching to a DOM configuration where animate objects are preferentially marked; furthermore, the timecourse with which this differential marking develops suggests this may be influenced both by participants own experience of the ambiguity of unmarked animate objects, and the requirements of their interlocutor. In sum, we see no evidence of a bias towards efficient communication in learning, but participants' behaviour during actual communicative interaction is clearly driven by communicative efficiency.

6. General Discussion

6.1. Other instances of communicative biases in learning

We think our data provides strong evidence against the efficient-communication-in-learning hypothesis advanced in FNJ. However, as mentioned briefly in the introduction, FNJ is not the only paper presenting experimental data from artificial language learning experiments which is consistent with biases in learning favouring communicatively-efficient or informative systems. What do our findings here imply for these other results, and the more general claim that learning biases may in some cases favour systems which are well-designed for communication? One possibility is simply that the biases for Differential Case Marking which FNJ report are particularly fragile, and that while this specific example is not robust, the more general case still stands, i.e. there are biases for communicative function in learning in at least some cases. However, given that this conclusion is at odds with the well-established literature on simplicity biases in learning (both in artificial language paradigms, e.g. Kirby et al., 2008; Silvey et al., 2015; Carr et al., 2017, 2020; Smith et al., 2020, and more generally, e.g. Chater & Vitanyi, 2003; Feldman, 2016), we think it is worth considering each piece of experimental evidence rather carefully.

Some putative examples of communicative biases in learning can be accounted for as misdiagnoses of simplicity biases, which can sometimes produce behaviours that are consistent with biases favouring informativeness. Carstensen et al. (2015) constitutes one such example. They argue that pressures for informativeness (i.e. favouring the reliable recovery of a signaller's intended meaning) might operate during learning, and show that category systems which are repeatedly learned and reproduced in a non-communicative learning-and-recall iterated learning design tend to become increasingly informative; in particular, category systems evolve from an initial random configuration to one where categories are contiguous, such that similar meanings fall in the same category. Contiguous categories are better for communication than non-contiguous categories, in that they direct the receiver of the category label to the right region of the semantic space, even if they fail to pick out exactly the right meaning. However, Carr et al. (2020) note that while simplicity and informativeness are often opposed (e.g. having few labels is simple but not informative), there are cases where the biases coincide: in particular, simplicity also favours contiguous categories, since contiguous categories can be represented more compactly and are therefore simpler. Carr et al. (2020) show that this can account for the results reported in Carstensen et al. (2015).

in other words, our failure to elicit reliable DOM effects in learning is unlikely to be due to an insufficiently clear contrast between animate and inanimate stimuli.

In particular, the apparent increase in informativeness occurring over generations of learning is likely to be driven by a simplicity-based preference for category contiguity which happens to also increase the notional communicative function of the evolving category systems. This may be a feature of the early stages of the evolution of category systems via learning, before categories collapse into one another.

Typicality matching or markedness matching might also account for some other putative cases of efficient communication biases in learning, for instance the asymmetries in number marking tackled by Kurumada & Grimm (2019). In English, singulars are unmarked whereas plurals are typically marked (e.g. *girl-*$\emptyset$ vs *girl-s*); however, in some languages this can be reversed for some nouns, with the singular being marked and the plural unmarked (e.g. *pys-en*, pea, vs *pys-*$\emptyset$, peas, in Welsh). Haspelmath & Karjus (2017) show that in five languages which show this marking reversal, the nouns which are linguistically marked in the singular but unmarked in the plural tend to be used to refer to multiplex concepts, i.e. groups of things. For example, while it is more common to talk about a single girl than multiple girls, it is more common to talk about multiple peas than a single pea. This configuration is of course optimal from the perspective of communicative efficiency: the system encodes number only when its meaning cannot be inferred from world knowledge. Kurumada & Grimm (2019) provide experimental evidence for a bias favouring precisely this configuration in an artificial language learning experiment: participants overproduce marked plurals for uniplex nouns (i.e. which typically occur singly) and under-produce marked plurals for multiplex nouns (which typically occur in groups). While Kurumada & Grimm follow FNJ in attributing this bias to efficient communication principals, it could equally be attributed to iconicity biases (markedness matching or typicality matching), following the same logic we outline here: unusual semantics (a plural uniplex noun, a single multiplex noun) should be reflected in an unusual or weightier form.[27]

However, there are cases which are harder to provide alternative explanations for, and which therefore constitute stronger evidence for biases for efficient communication in learning. Languages seem to trade off syntactic and morphological complexity: languages with richer morphology tend to exhibit greater syntactic flexibility, whereas more impoverished morphology is associated with more restrictive syntactic constraints (e.g. Sinnemäki, 2008; Koplenig et al., 2017). This is also true for word order flexibility and casemarking, where languages with richer systems of casemarking tend to allow more flexible word order, and languages with fixed word order are less likely to mark case or have fewer case distinctions (Lester et al., 2018), presumably because these two linguistic devices overlap in their com-

[27]Levshina (2018) constitutes another such example, where participants in an artificial language paradigm show a preference to use slightly shorter forms (5 rather than 6 syllables) to describe frequent events and longer forms for infrequent events. Again, this could be due to efficient communication considerations, but iconicity preferences could also account for this pattern. It is worth noting that Kanwal et al. (2017b) provide a similar manipulation of frequency and word length in a lexical learning task (with a much greater difference in effort between short and long forms) and show no evidence of participants producing the more communicative-efficient configuration aligning frequency and word length in a learning-only task (but a clear effect in communicative tasks); furthermore, this result replicates when raw frequency is replaced with predictability in context (Kanwal et al., 2017a); participants only behave in a communicatively-efficient manner when engaged in a communicative task. This mismatches Levshina (2018) but aligns with our more general claim that biases in learning tend to be agnostic with respect to communication.

municative function: providing two mechanisms to convey the same information (i.e. fixing word order and still marking case) is therefore redundant and a violation of efficiency considerations. Fedzechkina et al. (2017) show that this typological generalization is mirrored in participants' biases in artificial language learning. They trained participants on languages with optional casemarking on objects[28], where word order was varied between subjects and was either variable (SOV and OSV order were equally frequent in the input) or fixed (SOV order only). They report two main results. Firstly, participants in the variable word order condition tended to produce variable word order and variable casemarking; however, they tended to condition casemarking on word order, with OSV order being more likely to be marked. This behaviour could be driven by processing or communicative considerations, but equally could be an instance of iconicity via markedness matching (where the unusual order is marked linguistically). The more striking result occurs in the fixed order condition, where participants tend to reduce their use of the case marker, as predicted under the efficient communication accounts. Furthermore, Fedzechkina & Jaeger (2020) show that this result only holds if producing case markers is effortful (i.e. requires extra mouse clicks for participants, rather than simply selecting a single fully-inflected noun); however, this cannot be explained away as merely a preference for minimising production effort, since case markers are retained in the variable word order conditions.

We believe this behaviour in the fixed word order condition constitutes the clearest evidence to date for biases favouring efficient communication in learning. We can however offer a candidate alternative explanation. There is good evidence that conditioning of variation in other artificial language learning experiments offers some insulation against the elimination of variation: for instance, Hudson Kam & Newport (2009) show that participants are better able to reproduce a variable marker if its use is lexically conditioned (i.e. some nouns occur with the marker and others don't) than if it occurs in free variation (i.e. where all nouns sometimes take the marker); in the latter case, participants tend to reduce variation, over-using the most frequent marker. Samara et al. (2017) provide a similar result for the acquisition of socially-conditioned variation. Smith & Wonnacott (2010) and Smith et al. (2017) show in an iterated learning design that artificial languages with variable plural marking (two possible ways of marking plurality) tend to evolve to one of two configurations, either zero variability (one way of marking plurality is lost) or stably conditioned variability (both ways of marking plurality are retained, but the choice of marker becomes lexically conditioned), again suggesting that conditioning of variation, i.e. making that variation dependent on some other variable element of the linguistic context, insulates a varying element against the tendency for variation to be lost in learning. It may be that the word order variability in the variable order conditions of Fedzechkina et al. (2017) provides a salient context on which casemarking can be conditioned, and that conditioning provides some protection from production effort considerations; the fixed order condition lacks this conditioning context (since there is no variation in word order) and casemarking tends to regularise, with the zero-marking outcome being preferred simply because it is less effort to drop the markers than to include them everywhere. However, this alternative explanation is quite speculative, and while it receives circumstantial support from the papers we mention above, it would

[28]NB all objects were animate, i.e. this experiment does not address Differential Object Marking.

benefit from direct support in paradigms more closely matched to those used in Fedzechkina et al. (2017) and Fedzechkina & Jaeger (2020), showing that conditioning offered this kind of protection even in non-functional cases (where the effortful markers did not also serve a potential disambiguation function).[29]

6.2. A preference to mark OSV word order

Across our experiments, we repeatedly find that participants selectively fail to learn that casemarking is conditioned on word order: participants reliably learn this conditioning when OSV order is casemarked more than SOV order (in Experiment 2, and in the 40% casemarking language in Experiment 3), but fail to learn it when SOV order is marked more than OSV order (in Experiment 1, and in the 60% casemarking language in Experiment 3). This result is consistent with data from Fedzechkina et al. (2017), who report that participants tend to spontaneously condition casemarking on word order, preferring to mark OSV order, providing converging evidence that participants prefer to mark OSV order somehow. This is itself probably a reflection of an iconicity bias in learning, specifically markedness matching, where the unusual order (OSV, which places the object before the subject) is marked. As we noted in developing the case-order correlation account, this preference to mark OSV order is problematic when it comes to identifying the cause of any preference to mark animate objects, since participants are more likely to use OSV order with animate objects, due to an independently-attested preference to mention human referents before others, e.g. Meir et al. (2017). This highlights the challenges of identifying efficient communication biases in learning, since (at least in the early stages of learning) participants' behaviour is likely to be influenced by frequency distributions in their input interacting with multiple biases in learning (at the very least a preference to mention humans before other entities and to mark unusual word orders, possibly in conjunction with typicality or markedness matching preferences); designing paradigms which cleanly separate all these factors is challenging. Here we have at least attempted to separate out frequency effects in the input (e.g. the frequency of casemarking, the majority word order, correlations between order and case), perhaps with mixed success; however, we would be extremely wary of studies which do not take account of these factors and yet draw strong conclusions about the source of biases in participants output, as FNJ do.

6.3. Puzzles posed by our own data

In the introduction we highlighted an odd feature of FNJ Experiment 1, namely that participants substantially under-produced object casemarking on day 2, and highlighted

[29]It is also worth noting, as an aside, that we saw no evidence for a trade-off between word order variability and casemarking in our data, either in recall or communication; participants who only used one word order were no less likely to casemark objects than participants who exhibited variation in their word order. We verified this statistically with a model based on our recall and interaction data from day 4, combining across all 3 experiments: we measured word order variability as entropy of word orders (participants who used a single word order have word order entropy 0, participants who use SOV and OSV orders equally frequently have word order entropy 1), and found no relationship between word order entropy and proportion of marked objects in either day 4 recall ($b = 0.04, SE = 0.04$) or interaction ($b = 0.005, SE = 0.03$). While our experiment was not designed to test this relationship between word order variability and casemarking, it suggests that eliciting this bias is at least somewhat dependent on the details of the learning task.

the mismatch with their Experiment 2, where subjects produced subject casemarking at the correct frequency even on day 2. It seems only fair to acknowledge that our data is also messy! We see a clear and consistent signal of efficient communication operating in communicative interaction, but our results for non-communicative recall tests are much less clear, particularly in Experiment 3 where there was no consistent pattern of results supporting any of the three hypotheses we considered. In one sense this supports our overall contention, in that even if there *are* biases in learning favouring efficient communication these are weak and hard to spot, and likely to have negligible effects when compared to the robust bias for efficient communication operating during actual communication. However, it could be that our results are driven too much by noise, despite our large (relative to FNJ) sample size, and that a larger, cleaner sample might provide more definitive evidence for efficient communication or typicality matching biases in learning.

There is one feature of our data we would highlight as being particularly puzzling, namely the apparent effect in Experiment 3 of event composition (Subjects Can vs Cannot Be Objects) on casemarking early in learning (i.e. on day 2): in this experiment (but not in Experiments 1–2), participants in the Subjects Cannot Be Objects condition were *more* likely to mark animate objects than participants in the Subjects Can Be Objects condition; this difference is greater in the 60% casemarking languages. This tendency to differentiate animate objects in the Subjects Cannot Be Objects condition of course makes no sense under efficient communication considerations (it is the reverse of the expected effect, since marking animate objects is not necessary in the Subjects Cannot Be Objects condition). However, we are unable to offer a coherent explanation as to why it would occur. One possibility is that the Subjects Can/Cannot Be Objects factor somehow influences the perceived typicality of animate objects. If so, this pattern of results might be a product of markedness matching (but not typicality matching, since it is impervious to our manipulation of casemarking frequency): in the Subjects Can Be Objects condition the animate subjects/objects occur with high frequency and need to be marked less, whereas in the Subjects Cannot Be Objects condition the animate objects are more atypical and should (under markedness matching) be marked. However, this is rather speculative, and we see no evidence for this pattern of results in Experiments 1–2.

6.4. Future directions

As we have highlighted above, we think it will be useful to explore whether data taken as providing evidence for biases in efficient communication (e.g. Fedzechkina et al., 2017; Fedzechkina & Jaeger, 2020; Levshina, 2018; Kurumada & Grimm, 2019) can instead be explained by iconicity preferences; we also think our rather complex results show that multiple biases can intersect to shape participants' behaviour in artificial language paradigms, and that it requires substantial work to disentangle these various factors, which future work exploring efficient communication biases in learning needs to be wary of.

Having said that, our results show that pressures for efficient communication have quite strong effects during communicative interaction in artificial language paradigms (as also shown by e.g. Kanwal et al., 2017b,a), even when participants interact with a simulated partner. There are several obvious follow-up studies which could use our method to explore the consequences of efficient communication for Differential Case Marking systems. One is

to verify that Differential Subject Marking (tackled in FNJ Experiment 2) develops during communicative interaction in the same way as Differential Object Marking: we expect it will. A second possibility is to verify our tentative finding that participants adapt to the fine-grained communicative requirements of their partner. In our combined analysis of the communication phase of our experiment, we found some evidence that participants' behaviour during interaction was initially determined by their own experience of the ambiguity of unmarked animates, but that they adapted to the requirements of their simulated partner; in particular, Smeeble found unmarked animate objects ambiguous even in the Subjects Cannot Be Objects condition, and participants showed a tendency to learn to mark those objects more often even though they themselves did not experience them as ambiguous. It would be useful to verify that this change in the use of case markers does not occur in interaction with a version of Smeeble who, like participants, does not find unmarked animate objects ambiguous; we expect that in those circumstances participants would not selectively mark animate objects.

Finally, it is worth returning to the alternative accounts of Differential Case Marking outlined in the introduction. One prominent account argues that such systems arise from competing pressures to avoid ambiguity and minimize effort. This is consistent with our participants' behaviour during communicative interaction with Smeeble, where they use casemarkers more in situations where ambiguity would result from their absence. However, as reviewed in the introduction, Differential Object Marking in natural languages has a slightly puzzling feature that, in most DOM systems, atypical objects are casemarked regardless of their in-the-moment ambiguity (recall the example of "the murderer murdered his victim" in Spanish, where "victim" is rather redundantly marked as the object). While this has been taken as evidence that DOM is not motivated by ambiguity avoidance, our results suggest that in principle DOM may at least initially arise via this mechanism during language use. Over time, languages may come to obligatorily mark the types of objects which often (but not necessarily always) pose ambiguity through a process of regularization; this would remove the burden from speakers who must otherwise decide on a case-by-case basis how ambiguous an unmarked object would be, and from learners who must otherwise work out that potential ambiguity conditions casemarking in their input. This hypothesis could be straightforwardly tested in our paradigm, for example by designing sets of events such that some unmarked animate object would not be ambiguous (e.g. because a certain subset of animate nouns only ever functions as objects). We already know from our Subjects Cannot Be Objects condition that participants are sensitive to this kind of lack of ambiguity, although it would be useful to verify that it applies if only some nouns have this feature. Assuming participants can pick up on the fact that some subset of object nouns do not need to be marked as objects to be understood as such, a second step would be to test how communicative interaction affects their behaviour on those nouns. In particular, do they mark them, thus maintaining consistency with the nouns that must be marked, or do they selectively drop casemarking when it doesn't introduce ambiguity (as we expect)? If the latter, how do learners respond to such a system of variable object marking — do they acquire it, or do they tend to regularise such that all animate object nouns must be marked regardless of ambiguity? Based on the existing literature on learners' response to variability we think that regularization is the more likely outcome, and that learners will tend to extend casemarking beyond the nouns where it is required for ambiguity avoidance; this could explain how a global and obligatory pattern of

DOM could emerge even though object marking is initially driven purely by in-the-moment ambiguity avoidance.

7. Conclusions

Natural languages are well-designed for efficient communication. Two possible explanations for this have been suggested in the literature. One possibility is that this is due to biases operating in learning, whereby learners prefer languages which permit efficient communication (as argued by e.g. Fedzechkina et al., 2012). The alternative is that biases in learning are agnostic with respect to communicative function, and that some other mechanism — i.e. adaptation during actual communicative language use — is required to explain this aspect of language design (as argued by e.g. Kirby et al., 2015; Carr et al., 2020; Smith et al., 2020). On the face of it, the latter explanation is hard to reconcile with data from artificial language learning experiments reported by Fedzechkina et al. (2012) and related papers, showing apparent biases for efficient communication in learning. Here, we probe these results by replicating and extending Fedzechkina et al.'s paradigm for exploring the learning of Differential Object Marking, a communicatively-efficient system where casemarking is only used where required to avoid ambiguity. Contrary to Fedzechkina et al. (2012), across 3 experiments we find little evidence for a bias in learning favouring communicatively-efficient Differential Object Marking: participants' behaviour is impervious to the ambiguity of unmarked objects, and modulated by statistical properties of their input, observations which are inconsistent with their hypothesis and better explained by biases early in learning favouring iconicity in the form-meaning mapping, e.g. whereby atypical meanings are associated with atypical forms. However, we find good evidence that participants' behaviour in actual communicative language use in interaction *are* driven by efficient communication considerations: in interaction participants exhibited the expected Differential Object Marking pattern, and their behaviour is modulated both by their own perceptions of ambiguity and their interlocutor's response to unmarked objects. Based on this finding, we suggest that languages adapt to be communicative efficient as a result of being used in communication, rather than due to biases in human learning favouring communicatively-efficient languages.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No. 681942 held by KS and No. 757643 held by JC, and from the Experimental Psychology Society's Small Grants Scheme. Thanks to Holly Branigan for providing stimuli for our pilot experiment, Sara Rolando for drawing the stimuli for Experiments 1–3 here, and Hanna Jarvinen for assisting with stimuli preparation for all experiments. Thanks to Olga Fehér, Mora Maldonado, and Alan Smith for help with examples in Hungarian, Spanish, and German.

References

Aissen, J. (2003). Differential Object Marking: Iconicity vs. Economy. *Natural Language and Linguistic Theory*, *21*, 435–483.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*.

Bickel, B., & Witzlack-Makarevich, A. (2008). Referential scales and case alignment: Reviewing the typological evidence. In M. Richards, & A. Malchukov (Eds.), *Scales* (pp. 1–37). Leipzig: Universität Leipzig.

Börstell, C. (2019). Differential object marking in sign languages. *Glossa: a journal of general linguistics*, *4*, 3. doi:10.5334/gjgl.780.

Bossong, G. (1991). Differential Object Marking in Romance and beyond. In D. Kibbee, & D. Wanner (Eds.), *New Analyses in Romance Linguistics* (pp. 143–170). Amsterdam: Benjamins.

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic coordination in dialogue. *Cognition*, *75*, B13–B25.

Carr, J. W., Smith, K., Cornish, H., & Kirby, S. (2017). The Cultural Evolution of Structured Languages in an Open-Ended, Continuous World. *Cognitive Science*, *41*, 892–923. doi:10.1111/cogs.12371.

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

Carstensen, A., Xu, J., Smith, C. T., & Regier, T. (2015). Language evolution in the lab tends toward informative communication. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Chater, N., & Vitanyi, P. (2003). Simplicity: A unifying principle in cognitive science. *Trends in Cognitive Sciences*, *7*, 19–22.

Comrie, B. (1989). *Language Universals and Linguistic Typology*. (2nd ed.). Oxford: Blackwell.

Culbertson, J. (2012). Typological Universals as Reflections of Biased Learning: Evidence from Artificial Language Learning. *Language and Linguistics Compass*, *6*, 310–329.

Culbertson, J. (forthcoming). Artificial language learning. In J. Sprouse (Ed.), *Oxford Handbook of Experimental Syntax*. Oxford: Oxford University Press.

Culbertson, J., & Adger, D. (2014). Language learners privilege structured meaning over surface frequency. *Proceedings of the National Academy of Sciences*, *111*, 5842–5847. doi:10.1073/pnas.1320525111.

Culbertson, J., & Kirby, S. (2016). Simplicity and Specificity in Language: Domain-General Biases Have Domain-Specific Effects. *Frontiers in Psychology*, *6*, 1964.

Culbertson, J., Smolensky, P., & Legendre, G. (2012). Learning biases predict a word order universal. *Cognition*, *122*, 306–329. doi:10.1016/j.cognition.2011.10.017.

de Hoop, H., & Malchukov, A. L. (2008). Case-Marking Strategies. *Linguistic Inquiry*, *39*, 565–587. doi:10.1162/ling.2008.39.4.565.

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, Iconicity, and Systematicity in Language. *Trends in Cognitive Sciences*, *19*, 603–615. doi:10.1016/j.tics.2015.07.013.

Dixon, R. M. W. (1979). Ergativity. *Language*, *55*, 59–138.

Fedzechkina, M., Chu, B., & Florian Jaeger, T. (2018). Human Information Processing Shapes Language Change. *Psychological Science*, *29*, 72–82. doi:10.1177/0956797617728726.

Fedzechkina, M., & Jaeger, T. F. (2020). Production efficiency can cause grammatical change: Learners deviate from the input to better balance efficiency against robust message transmission. *Cognition*, *196*, 104115. doi:10.1016/j.cognition.2019.104115.

Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2012). Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, *109*, 17897–17902. doi:10.1073/pnas.1215776109.

Fedzechkina, M., Newport, E. L., & Jaeger, T. F. (2017). Balancing Effort and Information Transmission During Language Acquisition: Evidence From Word Order and Case Marking. *Cognitive Science*, *41*, 416–446. doi:10.1111/cogs.12346.

Fehér, O., Ritt, N., & Smith, K. (2019). Asymmetric accommodation during interaction leads to the regularisation of linguistic variants. *Journal of Memory and Language*, *109*, 104036. doi:10.1016/j.jml.2019.104036.

Feldman, J. (2016). The simplicity principle in perception and cognition: The simplicity principle. *Wiley Interdisciplinary Reviews: Cognitive Science*, *7*, 330–340. doi:10.1002/wcs.1406.

Ferdinand, V., Kirby, S., & Smith, K. (2019). The cognitive roots of regularization in language. *Cognition*, *184*, 53–68.

Givón, T. (1991). Markedness in Grammar: Distributional, Communicative and Cognitive Correlates of Syntactic Structure. *Studies in Language*, *15*, 335–370. doi:10.1075/sl.15.2.05giv.

Hahn, M., Jurafsky, D., & Futrell, R. (2020). Universals of word order reflect optimization of grammars for efficient communication. *Proceedings of the National Academy of Sciences*, *117*, 2347–2353. doi:10.1073/pnas.1910923117.

Haspelmath, M. (2008). Frequency vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics*, *19*, 1–33. doi:10.1515/COG.2008.001.

Haspelmath, M. (2018). No progress on differential object marking: Comments and reflections on Kalin (2018).

Haspelmath, M. (forthcoming). Role-reference associations and the explanation of argument coding splits. *Linguistics*, .

Haspelmath, M., & Karjus, A. (2017). Explaining asymmetries in number marking: Singulatives, pluratives, and usage frequency. *Linguistics*, *55*. doi:10.1515/ling-2017-0026.

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, *1*, 151–195.

Hudson Kam, C. L., & Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, *59*, 30–66. doi:10.1016/j.cogpsych.2009.01.001.

Imai, M., Kita, S., Nagumo, M., & Okada, H. (2008). Sound symbolism facilitates early verb learning. *Cognition*, *109*, 54–65. doi:10.1016/j.cognition.2008.07.015.

Jäger, G. (2007). Evolutionary Game Theory and Typology. A Case Study. *Language*, *83*, 74–109.

Kanwal, J., Smith, K., Culbertson, J., & Kirby, S. (2017a). Language-users choose short words in predictive contexts in an artificial language task. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 643–648). Austin, TX: Cognitive Science Society.

Kanwal, J., Smith, K., Culbertson, J., & Kirby, S. (2017b). Zipf's Law of Abbreviation and the Principle of Least Effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, *165*, 45–52. doi:10.1016/j.cognition.2017.05.001.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*, 1049–1054.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative Cultural Evolution in the Laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences, USA*, *105*, 10681–10686.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and Communication in the Cultural Evolution of Linguistic Structure. *Cognition*, *141*, 87–102.

Kittilä, S. (2005). Optional marking of arguments. *Language Sciences*, *27*, 483–514. doi:10.1016/j.langsci.2004.10.005.

Koplenig, A., Meyer, P., Wolfer, S., & Müller-Spitzer, C. (2017). The statistical trade-off between word order and word structure – Large-scale evidence for the principle of least effort. *PLOS ONE*, *12*, e0173614. doi:10.1371/journal.pone.0173614.

Kurumada, C., & Grimm, S. (2019). Predictability of meaning in grammatical encoding: Optional plural marking. *Cognition*, *191*, 103953. doi:10.1016/j.cognition.2019.04.022.

Lehmann, C. (1985). Grammaticalization: Synchronic variation and diachronic change. *Lingua e Stile*, *20*, 303–318.

Lester, N. A., Auderset, S., & Rogers, P. G. (2018). Case inflection and the functional indeterminacy of nouns:A cross-linguistic analysis. In C. Kalish, M. Rau, J. Zhu, & T. Rogers (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 2029–2034). Austin, TX: Cognitive Science Society.

Levshina, N. (2018). Linguistic Frankenstein, or How to test universal constraints without real languages. In K. Schmidtke-Bode, N. Levshina, S. M. Michaelis, & I. A. Seržant (Eds.), *Explanation in Typology: Diachronic Sources, Functional Motivations and the Nature of the Evidence* (pp. 205–222). Berlin: Language Sciences Press.

Martin, A., & White, J. (2019). Vowel Harmony and Disharmony Are Not Equivalent in Learning. *Linguistic Inquiry*, (pp. 1–20). doi:10.1162/ling_a_00375.

Meir, I., Aronoff, M., Börstell, C., Hwang, S.-O., Ilkbasaran, D., Kastner, I., Lepic, R., Ben-Basat, A. L., Padden, C., & Sandler, W. (2017). The effect of being human and the basis of grammatical word order: Insights from novel communication systems and young sign languages. *Cognition*, *158*, 189–207.

Moreton, E., & Pater, J. (2012). Structure and Substance in Artificial-phonology Learning, Part I: Structure: Structure and Substance in Artificial-Phonology Learning, Part I. *Language and Linguistics Compass*, *6*, 686–701. doi:10.1002/lnc3.363.

Nielsen, A. K., & Dingemanse, M. (2020). Iconicity in Word Learning and Beyond: A Critical Review. *Language and Speech*, (p. 002383092091433). doi:10.1177/0023830920914339.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*, 280–291.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, *111*, 317–328.

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, *104*, 1436–1441. doi:10.1073/pnas.0610341104.

Roberts, G., & Fedzechkina, M. (2018). Social biases modulate the loss of redundant forms in the cultural evolution of language. *Cognition*, *171*, 194–201. doi:10.1016/j.cognition.2017.11.005.

Saldana, C., Oseki, Y., & Culbertson, J. (2019). Do cross-linguistic patterns of morpheme order reflect a cognitive bias? In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Samara, A., Smith, K., Brown, H., & Wonnacott, E. (2017). Acquiring variation in an artificial language: Children and adults are sensitive to socially conditioned linguistic variation. *Cognitive Psychology*, *94*, 85–114. doi:10.1016/j.cogpsych.2017.02.004.

Schouwstra, M., & de Swart, H. (2014). The semantic origins of word order. *Cognition*, *131*, 431–436. doi:10.1016/j.cognition.2014.03.004.

Seržant, I. A. (2018). Weak universal forces: The discriminatory function of case in differential object marking systems. In K. Schmidtke-Bode, N. Levshina, S. M. Michaelis, & I. A. Seržant (Eds.), *Explanation in Typology: Diachronic Sources, Functional Motivations and the Nature of the Evidence* (pp. 151–179). Berlin: Language Science Press.

Silverstein, M. (1976). Hierarchy of features and ergativity. In R. M. W. Dixon (Ed.), *Grammatical Categories in Australian Languages* (pp. 112–171). New Jersey: Humanities Press.

Silvey, C., Kirby, S., & Smith, K. (2015). Word Meanings Evolve to Selectively Preserve Distinctions on Salient Dimensions. *Cognitive Science*, *39*, 212–226. doi:10.1111/cogs.12150.

Sinnemäki, K. (2008). Complexity trade-offs in core argument marking. In M. Miestamo, K. Sinnemäki, & F. Karlsson (Eds.), *Studies in Language Companion Series* (pp. 67–88). Amsterdam: John Benjamins Publishing Company volume 94. doi:10.1075/slcs.94.06sin.

Sinnemäki, K. (2014). A typological perspective on Differential Object Marking. *Linguistics*, *52*. doi:10.1515/ling-2013-0063.

Smith, K. (2004). The evolution of vocabulary. *Journal of Theoretical Biology*, *228*, 127–142. doi:10.1016/j.jtbi.2003.12.016.

Smith, K., Frank, S., Rolando, S., Kirby, S., & Loy, J. (2020). Simple kinship systems are more learnable. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*. Toronto: Cognitive Science Society.

Smith, K., Perfors, A., Fehér, O., Samara, A., Swoboda, K., & Wonnacott, E. (2017). Language learning, language use, and the evolution of linguistic variation. *Philosophical Transactions of the Royal Society B*, *372*, 20160051.

Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, *116*, 444–449.

St. Clair, M. C., Monaghan, P., & Ramscar, M. (2009). Relationships Between Language Structure and Language Learning: The Suffixing Preference and Grammatical Categorization. *Cognitive Science*, *33*, 1317–1329. doi:10.1111/j.1551-6709.2009.01065.x.

Thompson, B., Kirby, S., & Smith, K. (2016). Culture shapes the evolution of cognition. *Proceedings of the National Academy of Sciences*, *113*, 4530–4535. doi:10.1073/pnas.1523631113.

Thompson, R. L., Vinson, D. P., Woll, B., & Vigliocco, G. (2012). The Road to Language Learning Is Iconic: Evidence From British Sign Language. *Psychological Science*, *23*, 1443–1448. doi:10.1177/0956797612459763.

Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition*, *107*, 729–742. doi:10.1016/j.cognition.2007.08.007.

Wagner, S., Smith, K., & Culbertson, J. (2019). Acquiring agglutinating and fusional languages can be similarly difficult: Evidence from an adaptive tracking study. In *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (pp. 3050–3056). Montreal, QB: Cognitive Science Society.

White, J. (2014). Evidence for a learning bias against saltatory phonological alternations. *Cognition*, *130*, 96–115. doi:10.1016/j.cognition.2013.09.008.

Winters, J., Kirby, S., & Smith, K. (2015). Languages adapt to their contextual niche. *Language and Cognition*, *7*, 415–449. doi:10.1017/langcog.2014.35.

Winters, J., Kirby, S., & Smith, K. (2018). Contextual predictability shapes signal autonomy. *Cognition*, *176*, 15–30. doi:10.1016/j.cognition.2018.03.002.

Witzlack-Makarevich, A., & Seržant, I. A. (2018). Differential Argument Marking: Patterns Of Variation. In I. A. Seržant, & A. Witzlack-Makarevich (Eds.), *The Diachronic Typology of Differential Argument Marking* (pp. 1–40). Berlin: Language Science Press. doi:10.5281/ZENODO.1228243.

Zipf, G. K. (1936). *The Psycho-Biology of Language*. London: Routledge.

Zipf, G. K. (1949). *Human Behaviour and the Principle of Least Effort : An Introduction to Human Ecology*. Cambridge, MA: Addison-Wesley.

A. Summary tables for statistical analyses

Table 1: Summary table for the statistical analysis of identification accuracy in sentence comprehension trials for Experiment 1 (animate objects only). We ran a logit regression predicting accuracy (correct or incorrect) based on fixed effects of Day, Casemarking, Event Composition and their interaction. The model includes by-participant random intercepts only; models with random slopes for Casemarking and/or Day produced singular fits. Casemarking, Day and Event Composition are treatment coded: the model intercept therefore reflects log-odds of correctly identifying the subject on day 1, where the object is not casemarked, in the Subjects Cannot Be Objects condition. Note: $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

	b	SE	p
Intercept (unmarked, Day=1, Subjects Cannot Be Objects)	0.49	0.11	$< .001^{***}$
Casemarking	1.40	0.10	$< .001^{***}$
Day=2	0.85	0.11	$< .001^{***}$
Day=3	1.60	0.13	$< .001^{***}$
Day=4	2.24	0.16	$< .001^{***}$
Subjects Can Be Objects	-0.57	0.14	$< .001^{***}$
Casemarking * Day=2	0.40	0.18	$.023^{*}$
Casemarking * Day=3	0.04	0.21	.846
Casemarking * Day=4	0.09	0.27	.754
Subjects Can Be Objects * casemarking	0.19	0.13	.139
Subjects Can Be Objects * Day=2	-0.92	0.14	$< .001^{***}$
Subjects Can Be Objects * Day=3	-1.52	0.16	$< .001^{***}$
Subjects Can Be Objects * Day=4	-2.26	0.18	$< .001^{***}$
Subjects Can Be Objects * casemarking * Day=2	0.37	0.22	.099
Subjects Can Be Objects * casemarking * Day=3	1.28	0.26	$< .001^{***}$
Subjects Can Be Objects * casemarking * Day=4	1.63	0.32	$< .001^{***}$

Table 2: Summary table for the statistical analysis of casemarking in Experiment 1, during the sentence production test. We ran a logit regression predicting marking (marked or unmarked) based on fixed effects of Day, Animacy of object, Event Composition and their interaction, with by-participant random intercepts and by-participant random slopes for Day, Animacy, and their interaction. Day was treatment-coded, with Day 2 as the reference level. Animacy and Event Composition were deviation-coded; the model intercept therefore reflects the probability of casemarking an object at Day 2 collapsing across Animacy and Event Composition, and positive effects reflect more casemarking for animate objects / in the Subjects Can Be Objects condition.

	b	SE	p
Intercept (Day=2)	0.50	0.28	.079
Animacy	-0.43	0.22	.047*
Day=3	0.27	0.19	.157
Day=4	0.46	0.25	.067
Event Composition	-0.15	0.56	.793
Animacy * Day=3	0.15	0.24	.518
Animacy * Day=4	0.69	0.23	.002***
Animacy * Event Composition	0.01	0.43	.980
Event Composition * Day=3	0.53	0.39	.168
Event Composition * Day=4	0.60	0.50	.227
Animacy * Event Composition * Day=3	-0.21	0.48	.655
Animacy * Event Composition * Day=4	-0.06	0.45	.896

Table 3: Summary table for the statistical analysis of casemarking on day 4 in Experiment 1, during the sentence production test (Recall) and interaction with Smeeble (Interaction). We ran a logit regression predicting casemarking based on fixed effects of Block (Recall or Interaction), Animacy of object, Event Composition and their interactions, with by-participant random intercepts and by-participant random slopes for Block, Animacy, and their interaction. Block was treatment coded, Animacy and Event Composition were deviation-coded: the model intercept therefore reflects the probability of casemarking during pre-interaction Recall, collapsing across Animacy and Event Composition, and positive effects of Animacy / Event Composition reflect greater use of casemarking for animate objects / in the Subjects Can Be Objects condition.

	b	SE	p
Intercept (Block=Recall)	0.87	0.33	.007**
Animacy	0.19	0.21	.363
Block = Interaction	0.95	0.22	< .001***
Event Composition	0.53	0.65	.417
Animacy * Block=Interaction	0.43	0.20	.033*
Animacy * Event Composition	-0.01	0.42	.990
Event Composition * Block=Interaction	0.12	0.44	.778
Animacy * Event Composition * Block=Interaction	0.22	0.37	.547

Table 4: Summary table for the statistical analysis of word order in Experiment 1, during the sentence production test. We ran a logit regression predicting use of SOV word order based on fixed effects of Day, Animacy, Event Composition and their interactions, with by-participant random intercepts and by-participant random slopes for Day and Animacy (the model including random slope for the Day x Animacy interaction produced a singular fit). We used the same coding scheme as used in the analysis of casemarking: the model intercept reflects the probability of SOV order at Day 2 collapsing across Animacy and Event Composition, and positive effects of Animacy / Event Composition reflect greater use of SOV for animate objects / in the Subjects Can Be Objects condition.

	b	SE	p
Intercept (Day=2)	2.39	0.26	$< .001^{***}$
Animacy	-0.93	0.17	$< .001^{***}$
Day=3	-0.35	0.22	.108
Day=4	-0.28	0.25	.266
Event Composition	-0.39	0.50	.435
Animacy * Day=3	0.55	0.14	$< .001^{***}$
Animacy * Day=4	0.33	0.15	$.028^{*}$
Animacy * Event Composition	-0.07	0.31	.825
Event Composition * Day=3	-0.31	0.41	.446
Event Composition * Day=4	0.17	0.47	.723
Animacy * Event Composition * Day=3	-0.23	0.28	.422
Animacy * Event Composition * Day=4	0.09	0.29	.762

Table 5: Summary table for the statistical analysis of the effect of word order on casemarking in Experiment 1. We ran a logit regression predicting use of casemarking based on fixed effects of Day, Word Order and their interactions, with by-participant random intercepts and by-participant random slopes for Day, Word Order, and their interaction. Day was treatment-coded, and Word Order was deviation coded: the model intercept reflects the probability of casemarking at Day 2 collapsing across the two word orders, and positive effects of Word Order indicate greater use of casemarking for SOV sentences.

	b	SE	p
Intercept (Day=2)	0.44	0.26	.094
Word Order (at Day=2)	0.19	0.33	.567
Day=3	0.16	0.19	.412
Day=4	0.38	0.24	.104
Word Order * Day=3	0.06	0.37	.882
Word Order * Day=4	-0.10	0.39	.790

Table 6: Summary table for the statistical analysis of the effect of animacy and word order on casemarking in Experiment 1. We ran a logit regression predicting use of casemarking based on fixed effects of Day, Animacy, Event Composition, Word Order and their interactions, pulling out interactions between Event Composition and Word Order which we felt were not meaningful. The random effects included by-participant random intercepts and by-participant random slopes for Day, Animacy, Word Order, and the Day x Word Order interaction; including any other interactions in the random slopes caused singular fit. We used the same coding scheme as used in the analysis of casemarking (see Table 2): the model intercept reflects the probability of casemarking at Day 2 collapsing across Animacy and Word Order, and positive effects of Word Order / Animacy reflect greater use of casemarking for SOV sentences / animate objects.

	b	SE	p
Intercept (Day=2)	0.55	0.29	.058
Animacy	-0.59	0.22	.007**
Word Order	0.07	0.36	.841
Day=3	0.10	0.20	.629
Day=4	0.39	0.26	.134
Event Composition	-0.02	0.60	.975
Animacy * Day = 3	0.07	0.18	.679
Animacy * Day = 4	0.65	0.19	< .001***
Word Order * Day=3	0.26	0.41	.531
Word Order * Day=4	-0.01	0.42	.988
Animacy * Word Order	0.69	0.35	.048*
Animacy * Event Composition	-0.20	0.41	.616
Event Composition * Day=3	0.30	0.39	.437
Event Composition * Day=4	0.21	0.52	.684
Animacy * Word Order * Day=3	0.50	0.36	.167
Animacy * Word Order * Day=4	-0.12	0.37	.739
Animacy * Event Composition * Day=3	-0.04	0.31	.902
Animacy * Event Composition * Day=4	-0.05	0.32	.878

Table 7: Summary table for the statistical analysis of identification accuracy in sentence comprehension trials for Experiment 2 (animate objects only). We ran a logit regression predicting accuracy based on fixed effects of Day, Casemarking, Event Composition and their interactions. The random effect structure included by-participant random intercepts and by-participant slopes for Casemarking; adding Day resulted in a singular fit. As in the equivalent analysis for Experiment 1, Casemarking, Day and Event Composition are again treatment coded: the model intercept therefore reflects log-odds of correctly identifying the subject on day 1, where the object is not casemarked, in the Subjects Cannot Be Objects condition.

	b	SE	p
Intercept (unmarked, Day=1, Subjects Cannot Be Objects)	0.98	0.12	$< .001^{***}$
Casemarking	0.86	0.20	$< .001^{***}$
Day=2	0.61	0.11	$< .001^{***}$
Day=3	1.29	0.12	$< .001^{***}$
Day=4	1.53	0.13	$< .001^{***}$
Subjects Can Be Objects	-0.73	0.16	$< .001^{***}$
Casemarking * Day=2	0.65	0.21	$.002^{**}$
Casemarking * Day=3	1.34	0.30	$< .001^{***}$
Casemarking * Day=4	1.58	0.36	$< .001^{***}$
Subjects Can Be Objects * casemarking	0.78	0.28	$.005^{**}$
Subjects Can Be Objects * Day=2	-0.60	0.14	$< .001^{***}$
Subjects Can Be Objects * Day=3	-1.27	0.15	$< .001^{***}$
Subjects Can Be Objects * Day=4	-1.47	0.16	$< .001^{***}$
Subjects Can Be Objects * casemarking * Day=2	-0.03	0.28	.905
Subjects Can Be Objects * casemarking * Day=3	-0.44	0.35	.212
Subjects Can Be Objects * casemarking * Day=4	-0.39	0.42	.356

Table 8: Summary table for the statistical analysis of the effect of animacy on word order in Experiment 2, during the sentence production test. We ran a logit regression predicting use of SOV word order based on fixed effects of Day, Animacy, and their interaction, with by-participant random intercepts and by-participant random slopes for Day and Animacy (the model including random slope for the Day x Animacy interaction produced a singular fit). We used the same coding scheme as used elsewhere: the model intercept reflects the probability of SOV order at Day 2 collapsing across Animacy, and a negative effect of Animacy reflects greater use of OSV for animate objects.

	b	SE	p
Intercept (Day=2)	2.86	0.27	$< .001^{***}$
Animacy	-1.28	0.21	$< .001^{***}$
Day=3	-0.25	0.20	.227
Day=4	-0.82	0.25	$.001^{**}$
Animacy * Day=3	0.35	0.17	$.044*$
Animacy * Day=4	0.47	0.17	$.006^{**}$

Table 9: Summary table for the statistical analysis of the effect of word order on casemarking in Experiment 2. We ran a logit regression predicting use of casemarking based on fixed effects of Day, Word Order and their interactions, with by-participant random intercepts and by-participant random slopes for Day, Word Order, and their interaction. Day was treatment-coded, and Word Order was deviation coded: the model intercept reflects the probability of casemarking at Day 2 collapsing across the two word orders, and positive effects of Word Order would indicate greater use of casemarking for SOV sentences (recall that in the input SOV sentences were casemarked *less*, so we expect a negative effect here).

	b	SE	p
Intercept (Day=2)	-1.17	0.33	$< .001^{***}$
Word Order (at Day=2)	-1.84	0.42	$< .001^{***}$
Day=3	0.42	0.24	.079
Day=4	0.24	0.26	.371
Word Order * Day=3	-0.40	0.44	.361
Word Order * Day=4	-0.37	0.47	.424

Table 10: Summary table for the statistical analysis of casemarking in Experiment 2, during the sentence production test. We ran a logit regression with the same structure as in Experiment 1, predicting marking based on fixed effects of Day, Animacy, Event Composition and their interactions. The model included by-participant random intercepts and by-participant random slopes for Day and Animacy; including their interaction resulted in a singular fit. Given the coding scheme used, the model intercept reflects the probability of casemarking an object at Day 2 collapsing across Animacy and Event Composition, positive effects for Animacy / Event Composition reflect more casemarking for animate objects / in the Subjects Can Be Objects condition.

	b	SE	p
Intercept (Day=2)	-1.85	0.29	$< .001^{***}$
Animacy	0.38	0.22	.074
Day=3	0.42	0.21	.052
Day=4	0.30	0.25	.223
Event Composition	-0.25	0.57	.657
Animacy * Day=3	-0.10	0.15	.514
Animacy * Day=4	-0.14	0.16	.357
Animacy * Event Composition	0.30	0.42	.465
Event Composition * Day=3	-0.11	0.40	.792
Event Composition * Day=4	0.84	0.47	.072
Animacy * Event Composition * Day=3	-0.32	0.30	.284
Animacy * Event Composition * Day=4	-0.52	0.31	.097

Table 11: Summary table for the statistical analysis of casemarking in Experiment 2, during the sentence production test, including Word Order (i.e. whether the participant used SOV or OSV order on a given trial) as an additional predictor. Since we did had no reason to predict word order to interact with event composition the model structure was Animacy * Day * (Event Composition + Word Order); the random effects include by-participant random intercepts and slopes for all within-subject predictors (Animacy, Day, Word Order) and their interactions. Coding of fixed effects are as for those in Table 10; Word Order was deviation-coded such that the model intercept collapses across both orders and positive effects for order indicate more casemarking for SOV sentences. Note that since animacy influences participants' word order choice, these predictors are somewhat colinear, leading to several Variance Inflation Factors above 3 in this model.

	b	SE	p
Intercept (Day=2)	-1.22	0.35	$< .001^{***}$
Animacy	-0.14	0.25	.593
Word Order	-1.95	0.43	$< .001^{***}$
Day=3	0.38	0.26	.134
Day=4	0.25	0.28	.386
Event Composition	-0.49	0.58	.396
Animacy * Day = 3	0.42	0.29	.152
Animacy * Day = 4	0.17	0.31	.581
Word Order * Day=3	-0.36	0.44	.411
Word Order * Day=4	-0.44	0.50	.378
Animacy * Word Order	0.49	0.37	.192
Animacy * Event Composition	0.08	0.39	.843
Event Composition * Day=3	0.01	0.42	.987
Event Composition * Day=4	0.78	0.48	.103
Animacy * Word Order * Day=3	-0.85	0.47	.070
Animacy * Word Order * Day=4	-0.44	0.48	.357
Animacy * Event Composition * Day=3	-0.12	0.42	.772
Animacy * Event Composition * Day=4	-0.28	0.46	.547

Table 12: Summary table for the statistical analysis of casemarking in Experiment 1 and 2 combined, during the sentence production test. We ran a logit regression with the same structure as in Experiment 2 but with an additional fixed effect of Experiment, predicting marking based on fixed effects of Day, Animacy, Event Composition, Experiment and their interaction. Experiment is deviation-coded: given the coding scheme used, the model intercept reflects the probability of casemarking an object at Day 2 collapsing across Expereiment, Animacy and Event Composition, and positive effects for Experiment / Animacy / Event Composition reflect more casemarking in Experiment 2 / for animate objects / in the Subjects Can Be Objects condition.

	b	SE	p
Intercept (Day=2)	-0.70	0.20	$< .001^{***}$
Experiment	-2.36	0.40	$< .001^{***}$
Animacy	-0.06	0.15	.692
Event Composition	-0.22	0.40	.581
Day=3	0.30	0.14	$.027^{*}$
Day=4	0.37	0.17	$.030^{*}$
Experiment * Animacy	0.72	0.30	$.017^{*}$
Experiment * Event Composition	0.15	0.79	.850
Animacy * Event Composition	0.04	0.29	.881
Experiment * Day=3	0.03	0.28	.926
Experiment * Day=4	-0.20	0.35	.554
Animacy * Day=3	0.18	0.17	.298
Animacy * Day=4	0.35	0.17	$.037^{*}$
Event Composition * Day=3	0.022	0.27	.413
Event Composition * Day=4	0.75	0.34	$.027^{*}$
Experiment * Animacy * Event Composition	0.05	0.58	.926
Experiment * Animacy * Day=3	0.04	0.35	.907
Experiment * Animacy * Day=4	-0.67	0.34	.053
Experiment * Event Composition * Day=3	-0.53	0.53	.322
Experiment * Event Composition * Day=4	0.39	0.67	.561
Animacy * Event Composition * Day=3	-0.17	0.33	.617
Animacy * Event Composition * Day=4	-0.18	0.32	.563
Experiment * Animacy * Event Composition * Day=3	-0.02	0.66	.974
Experiment * Animacy * Event Composition * Day=4	-0.29	0.63	.642

Table 13: Summary table for the statistical analysis of casemarking on day 4 in Experiment 2, during the sentence production test (Recall) and interaction with Smeeble (Interaction). We ran a logit regression predicting casemarking based on fixed effects of Block (Recall or Interaction), Animacy, Event Composition and their interactions, with by-participant random intercepts and by-participant random slopes for Block, Animacy, and their interaction. Block was treatment coded, Animacy and Event Composition were deviation-coded: the model intercept therefore reflects the probability of casemarking during pre-interaction Recall, collapsing across Animacy and Event Composition, and positive effects of Animacy / Event Composition reflect greater use of casemarking for animate objects / in the Subjects Can Be Objects condition.

	b	SE	p
Intercept (Block=Recall)	-1.57	0.30	$< .001^{***}$
Animacy	0.30	0.27	.270
Block = Interaction	0.36	0.19	.056
Event Composition	0.44	0.57	.439
Animacy * Block=Interaction	0.98	0.25	$< .001^{***}$
Animacy * Event Composition	-0.24	0.49	.620
Event Composition * Block=Interaction	0.35	0.35	.326
Animacy * Event Composition * Block=Interaction	1.07	0.41	$.009^{**}$

Table 14: Summary table for the statistical analysis of identification accuracy in sentence comprehension trials for Experiment 3 (animate objects only). We analysed data from unmarked trials only, in order to simplify the statistic and since this is where the relevant ambiguity occurs; we ran a logit regression predicting accuracy based on fixed effects of Day, Event Composition, Proportion Marked (based on the participants' input language, 40% or 60%) and their interactions. The random effect structure included by-participant random intercepts and by-participant slopes for Day. As in the equivalent analysis for Experiment 1–2, Day and Event Composition are treatment coded, and Proportion Marked is deviation coded: the model intercept therefore reflects log-odds of correctly identifying the subject (from a description without casemarking) on day 1, in the Subjects Cannot Be Objects condition, averaging over the two input languages.

	b	SE	p
Intercept (unmarked, Day=1, Subjects Cannot Be Objects)	0.64	0.06	$< .001^{***}$
Day=2	0.99	0.11	$< .001^{***}$
Day=3	1.92	0.17	$< .001^{***}$
Day=4	2.71	0.23	$< .001^{***}$
Subjects Can Be Objects	-0.67	0.08	$< .001^{***}$
Proportion Marked	-0.23	0.11	$.039^{*}$
Subjects Can Be Objects * Day=2	-1.00	0.15	$< .001^{***}$
Subjects Can Be Objects * Day=3	-1.89	0.22	$< .001^{***}$
Subjects Can Be Objects * Day=4	-2.66	0.28	$< .001^{***}$
Proportion Marked * Day=2	0.01	0.21	.961
Proportion Marked * Day=3	-0.46	0.30	.121
Proportion Marked * Day=4	-0.91	0.36	$.012^{*}$
Subjects Can Be Objects * Proportion Marked	0.14	0.16	.360
Subjects Can Be Objects * Proportion Marked * Day=2	-0.03	0.28	.911
Subjects Can Be Objects * Proportion Marked * Day=3	0.31	0.40	.438
Subjects Can Be Objects * Proportion Marked * Day=4	0.85	0.47	.071

Table 15: Summary table for the statistical analysis of the effect of animacy on word order in Experiment 3, during the sentence production test. We ran a logit regression predicting use of SOV word order based on fixed effects of Day, Animacy, Proportion Marked, and their interaction, with by-participant random intercepts and by-participant random slopes for Day, Animacy and their interaction. We used the same coding scheme as used elsewhere: the model intercept reflects the probability of SOV order at Day 2 collapsing across Animacy and Proportion Marked, positive effects of Animacy / Proportion Marked reflects greater use of SOV for animate objects / in the 60% casemarking language.

	b	SE	p
Intercept (Day=2)	0.68	0.25	.007**
Animacy	-0.26	0.17	.121
Day=3	-0.71	0.18	< .001***
Day=4	-0.67	0.21	.001**
Proportion Marked	0.89	0.50	.075
Animacy * Day=3	0.26	0.18	.132
Animacy * Day=4	0.04	0.18	.822
Proportion Marked * Day=3	-0.31	0.36	.381
Proportion Marked * Day=4	-0.32	0.41	.428
Animacy * Proportion Marked	-0.06	0.32	.849
Animacy * Proportion Marked * Day=3	-0.01	0.34	.987
Animacy * Proportion Marked * Day=4	0.01	0.35	.972

Table 16: Summary table for the statistical analysis of the effect of word order on casemarking in Experiment 3. We ran a logit regression predicting use of casemarking based on fixed effects of Day, Word Order, Proportion Marked and their interactions, with by-participant random intercepts and by-participant random slopes for Day, Word Order, and their interaction. Day and Proportion Marked were treatment-coded, and Word Order was deviation coded: the model intercept reflects the probability of casemarking at Day 2 in the 40% casemarking language, collapsing across the two word orders, positive effects of Word Order / Proportion Marked indicate greater use of casemarking for SOV sentences / in the 60% casemarking condition.

	b	SE	p
Intercept (Day=2)	-1.53	0.32	< .001***
Word Order (at Day=2)	-0.10	0.28	.721
Day=3	0.34	0.21	.109
Day=4	0.46	0.26	.080
Proportion Marked	1.87	0.44	< .001***
Word Order * Day=3	0.21	0.32	.510
Word Order * Day=4	0.24	0.36	.512
Proportion Marked * Day=3	-0.54	0.30	.066
Proportion Marked * Day=4	-0.57	0.36	.120
Word Order * Proportion Marked	-1.49	0.39	< .001***
Word Order * Proportion Marked * Day=3	-0.25	0.44	.569
Word Order * Proportion Marked * Day=4	-0.70	0.50	.160

Table 17: Summary table for the statistical analysis of casemarking in Experiment 3, during the sentence production test. We ran a logit regression predicting marking based on fixed effects of Animacy, Day, Event Composition, Proportion Marked and their interactions. Day is treatment-coded, all other factors were deviation coded: given the coding scheme used, the model intercept reflects the probability of casemarking an object at Day 2 collapsing across Animacy, Event Composition, and Proportion Marked and positive effects for Animacy / Event Composition / Proportion Marked reflect more casemarking for animate objects / in the Subjects Can Be Objects condition / in the 60% casemarked input language.

	b	SE	p
Intercept (Day=2)	-0.75	0.22	$< .001^{***}$
Animacy	-0.07	0.17	.669
Event Composition	0.15	0.45	.741
Proportion Marked	2.02	0.45	$< .001^{***}$
Day=3	0.12	0.16	.458
Day=4	0.22	0.18	.236
Animacy * Event Composition	-0.67	0.32	$.036^{*}$
Animacy * Proportion Marked	-0.54	0.33	.102
Event Composition * Proportion Marked	2.48	0.88	$.005^{**}$
Animacy * Day=3	-0.09	0.18	.623
Animacy * Day=4	-0.09	0.18	.619
Event Composition * Day=3	-0.14	0.31	.658
Event Composition * Day=4	-0.10	0.36	.772
Proportion Marked * Day=3	-0.46	0.32	.143
Proportion Marked * Day=4	-0.48	0.37	.186
Animacy * Event Composition * Proportion Marked	-0.60	0.64	.352
Animacy * Event Composition * Day=3	0.47	0.33	.160
Animacy * Event Composition * Day=4	0.38	0.34	.269
Animacy * Proportion Marked * Day=3	0.16	0.35	.652
Animacy * Proportion Marked * Day=4	0.40	0.36	.273
Event Composition * Proportion Marked * Day=3	-0.31	0.62	.609
Event Composition * Proportion Marked * Day=4	-0.40	0.71	.572
Animacy * Event Composition * Proportion Marked * Day=3	1.06	0.67	.118
Animacy * Event Composition * Proportion Marked * Day=4	0.26	0.70	.713

Table 18: Summary table for the statistical analysis of casemarking on day 4 in Experiment 3, during the sentence production test (Recall) and interaction with Smeeble (Interaction). We ran a logit regression predicting casemarking based on fixed effects of Block (Recall or Interaction), Animacy, Event Composition, Proportion Marked, and their interactions, with by-participant random intercepts and by-participant random slopes for Block, Animacy, and their interaction. Block was treatment coded, Animacy, Event Composition and Proportion Marked were deviation-coded: the model intercept therefore reflects the probability of casemarking during pre-interaction Recall, collapsing across Animacy, Event Composition and Proportion Marked, and positive effects of Animacy / Event Composition / Proportion Marked reflect greater use of casemarking for animate objects / in the Subjects Can Be Objects condition / in the 60% casemarking input language.

	b	SE	p
Intercept (Block=Recall)	-0.50	0.24	$.035^{*}$
Animacy	-0.04	0.16	.808
Block = Interaction	0.62	0.16	$< .001^{***}$
Event Composition	-0.13	0.47	.783
Proportion Marked	1.17	0.47	$.013^{*}$
Animacy * Block=Interaction	0.61	0.17	$< .001^{***}$
Animacy * Event Composition	-0.33	0.30	.271
Block=Interaction * Event Composition	0.03	0.33	.930
Animacy * Proportion Marked	-0.18	0.30	.555
Block=Interaction * Proportion Marked	-0.09	0.33	.773
Event Composition * Proportion Marked	2.12	0.94	$.024^{*}$
Animacy * Block=Interaction * Event Composition	0.39	0.32	.228
Animacy * Block=Interaction * Proportion Marked	-0.53	0.33	.104
Animacy * Event Composition * Proportion Marked	-0.30	0.61	.626
Block=Interaction * Event Composition * Proportion Marked	0.64	0.65	.324
Animacy * Block=Interaction * Event Composition * Proportion Marked	1.08	0.66	.106

Table 19: Summary table for the statistical analysis of casemarking in Experiments 1– 3 combined, during the sentence production test. We ran a logit regression predicting marking based on fixed effects of Animacy, Day, Event Composition, Proportion Marked, Order Casemarked More (in the input: SOV or OSV) and their interactions. Day is treatment-coded, all other factors were deviation coded: given the coding scheme used, the model intercept reflects the probability of casemarking an object at Day 2 collapsing across Animacy, Event Composition, Proportion Marked and Order Casemarked More, and positive effects for Animacy / Event Composition / Proportion Marked / Order Casemarked More reflect more casemarking for animate objects / in the Subjects Can Be Objects condition / in the 60% casemarked input language / in input languages where SOV is casemarked more often than OSV.

	b	SE	p
Intercept (Day=2)	-0.72	0.15	$< .001^{***}$
Animacy	-0.07	0.11	.523
Proportion Marked	2.19	0.30	$< .001^{***}$
Order Marked More	0.17	0.30	.576
Event Composition	-0.04	0.30	.895
Day=3	0.21	0.10	$.043^{*}$
Day=4	0.30	0.13	$.018^{*}$
Animacy * Proportion Marked	-0.63	0.22	$.005^{**}$
Animacy * Order Marked More	-0.13	0.22	.557
Proportion Marked * Order Marked More	0.11	0.57	.855
Animacy * Event Composition	-0.31	0.22	.148
Proportion Marked * Event Composition	1.30	0.57	$.023^{*}$
Order Marked More * Event Composition	-1.16	0.57	$.043^{*}$
Animacy * Day=3	0.05	0.12	.708
Animacy * Day=4	0.14	0.12	.258
Proportion Marked * Day=3	-0.24	0.21	.251
Proportion Marked * Day=4	-0.14	0.25	.567
Order Marked More * Day=3	0.19	0.20	.342
Order Marked More * Day=4	0.36	0.25	.146
Event Composition * Day=3	0.06	0.20	.786
Event Composition * Day=4	0.33	0.25	.182
Animacy * Proportion Marked * Order Marked More	0.13	0.43	.764
Animacy * Proportion Marked * Event Composition	-0.28	0.43	.510
Animacy * Order Marked More * Event Composition	0.24	0.43	.577
Proportion Marked * Order Marked More * Event Composition	-0.73	1.09	.504
Animacy * Proportion Marked * Day=3	0.07	0.25	.788
Animacy * Proportion Marked * Day=4	0.54	0.25	$.029^{*}$
Animacy * Order Marked More * Day=3	-0.02	0.24	.928
Animacy * Order Marked More * Day=4	0.20	0.24	.397
Proportion Marked * Order Marked More * Day=3	0.38	0.40	.351
Proportion Marked * Order Marked More * Day=4	0.30	0.49	.541
Animacy * Event Composition * Day=3	0.16	0.24	.485
Animacy * Event Composition * Day=4	0.10	0.23	.681
Proportion Marked * Event Composition * Day=3	0.10	0.40	.801
Proportion Marked * Event Composition * Day=4	-0.36	0.49	.461
Order Marked More * Event Composition * Day=3	0.39	0.40	.331
Order Marked More * Event Composition * Day=4	-0.01	0.49	.978
Animacy * Proportion Marked * Order Marked More * Event Composition	1.41	0.82	.083
Animacy * Proportion Marked * Order Marked More * Day=3	0.37	0.46	.420
Animacy * Proportion Marked * Order Marked More * Day=4	0.76	0.46	.099
Animacy * Proportion Marked * Event Composition * Day=3	0.51	0.47	.273
Animacy * Proportion Marked * Event Composition * Day=4	0.23	0.46	.615
Animacy * Order Marked More * Event Composition * Day=3	-0.48	0.47	.303
Animacy * Order Marked More * Event Composition * Day=4	0.06	0.46	.892
Proportion Marked * Order Marked More * Event Composition * Day=3	0.72	0.76	.343
Proportion Marked * Order Marked More * Event Composition * Day=4	1.72	0.90	.057
Animacy * Proportion Marked * Order Marked More * Event Composition * Day=3	-1.30	0.89	.144

Table 20: Summary table for the statistical analysis of casemarking on day 4 in all 3 experiments, during the sentence production test (Recall) and interaction with Smeeble (Interaction). We ran a logit regression predicting casemarking based on fixed effects of Block (Recall or Interaction), Animacy, Event Composition, and their interactions, with by-participant random intercepts and by-participant random slopes for Block, Animacy, and their interaction. Block was treatment coded, Animacy and Event Composition were deviation-coded: the model intercept therefore reflects the probability of casemarking during pre-interaction Recall, collapsing across Animacy and Event Composition, and positive effects of Animacy / Event Composition reflect greater use of casemarking for animate objects / in the Subjects Can Be Objects condition.

	b	SE	p
Intercept (Block=Recall)	-0.36	0.17	.040*
Animacy	0.15	0.11	.188
Block = Interaction	0.66	0.11	< .001***
Event Composition	0.33	0.34	.332
Animacy * Block=Interaction	0.68	0.11	< .001***
Animacy * Event Composition	-0.21	0.22	.349
Block=Interaction * Event Composition	0.15	0.22	.486
Animacy * Block=Interaction * Event Composition	0.52	0.21	.016*

Table 21: Summary table for the statistical analysis of the timecourse of casemarking during interaction on day 4 in all 3 experiments. We ran a logit regression predicting casemarking based on fixed effects of Animacy, Event Composition, Trial Number and their interactions, with by-participant random intercepts and by-participant random slopes for Animacy, Trial Number, and their interaction. Event Composition was treatment coded, Animacy was deviation-coded, and Trial Number over the 40 trials was coded 0–39; the model intercept therefore reflects performance on trial 1 in the Subjects Cannot Be Objects condition, positive effects of Animacy / Event Composition reflect greater use of casemarking for animate objects / in the Subjects Can Be Objects condition, and positive effects involving Trial Number indicate the increment in casemarking for every additional trial.

	b	SE	p
Intercept (Subjects Cannot Be Objects, Trial Number = 1)	0.22	0.29	.441
Animacy	0.40	0.23	.082
Trial Number	0.005	0.004	.284
Subjects Can Be Objects	0.22	0.41	.584
Animacy * Trial Number	0.01	0.01	.054
Animacy * Subjects Can Be Objects	0.69	0.33	.039*
Trial Number * Subjects Can Be Objects	0.01	0.01	.035*
Animacy * Trial Number * Subjects Can Be Objects	-0.02	0.01	.035**

B. Additional figures

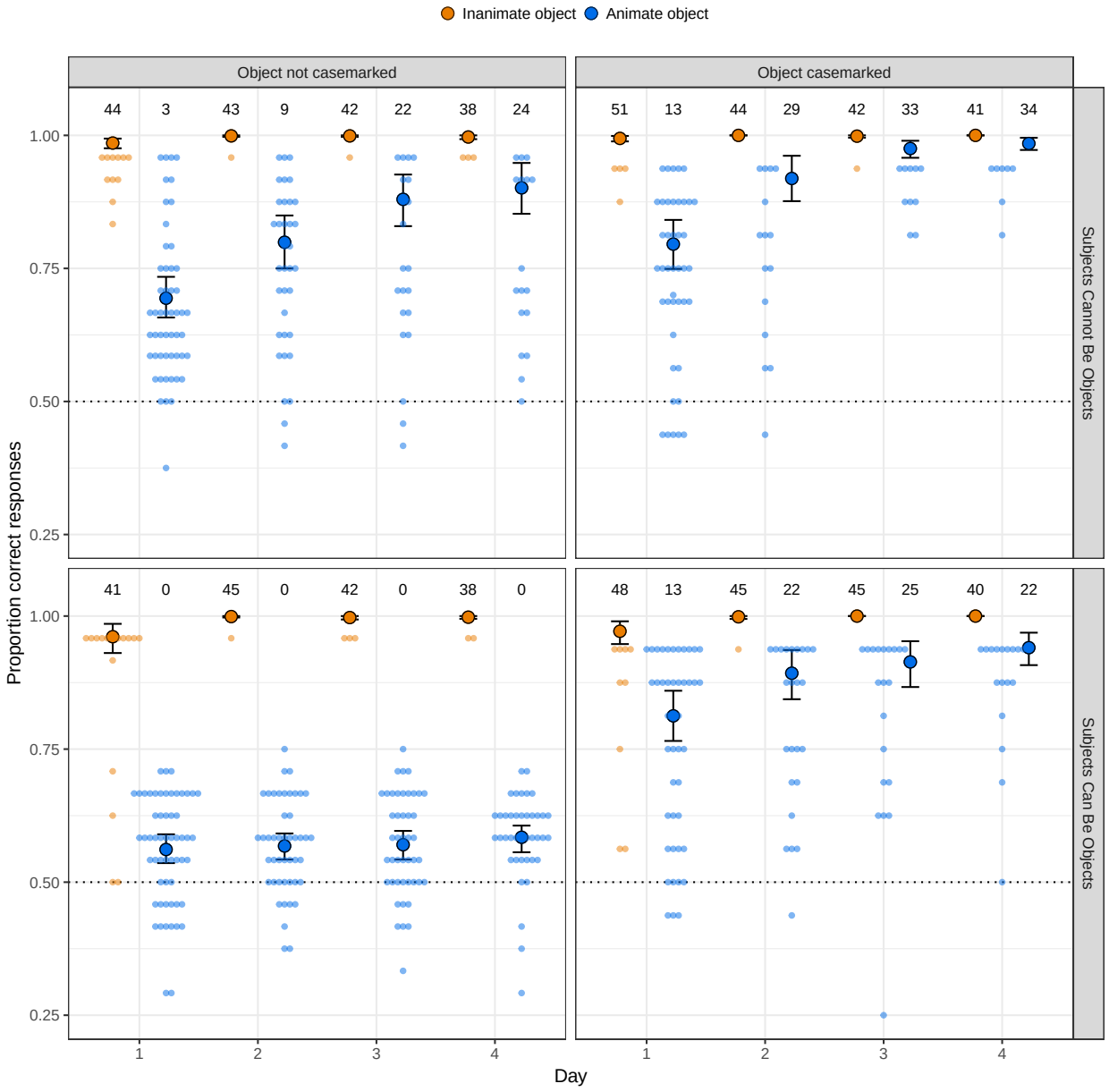

Figure 22: Proportion of correct responses on sentence comprehension trials in Experiment 2. There are too many participants performing at ceiling (i.e. all correct responses) to plot; instead, the text annotations give the number of participants producing ceiling performance. Performance is at ceiling when the object is inaninimate, and is high when the object is animate but casemarked; however, unmarked animate objects are only genuinely ambiguous in the Subjects Can Be Objects condition.

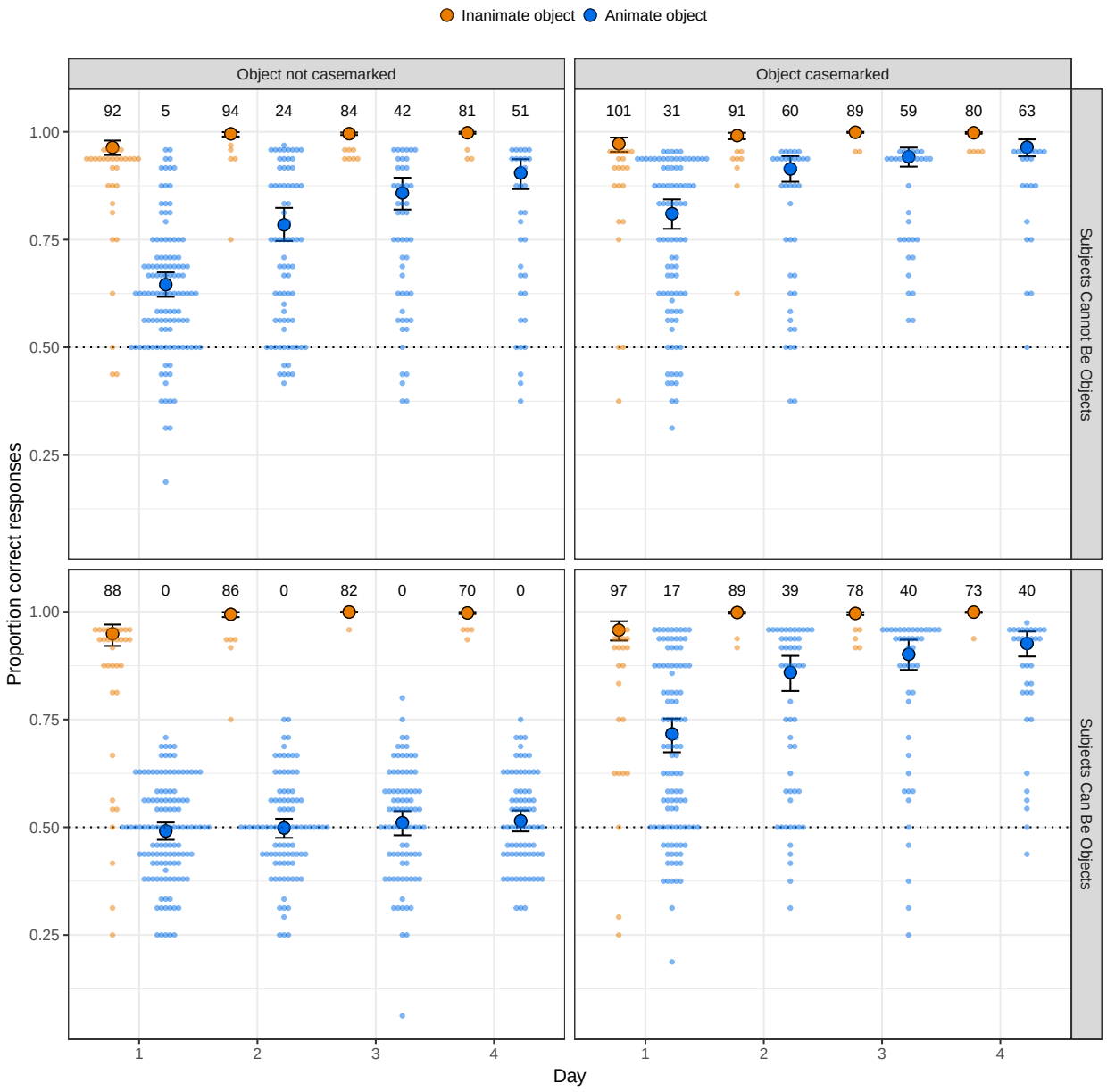

Figure 23: Proportion of correct responses on sentence comprehension trials in Experiment 3. There are too many participants performing at ceiling (i.e. all correct responses) to plot; instead, the text annotations give the number of participants producing ceiling performance. Performance is at ceiling when the object is inaninimate, and is high when the object is animate but casemarked; however, unmarked animate objects are only genuinely ambiguous in the Subjects Can Be Objects condition.