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

Kenny Smith, Jennifer Culbertson

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 the uncertainty or ambiguity experienced by a listener, despite the fact that they never use the artificial language in a communicative task where a listener's uncertainty is a relevant consideration. This is surprising given previous studies suggesting that biases in learning favour simplicity and are agnostic with respect to communicative function. We report an experiment 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' do not reliably produce Differential Object Marking in non-communicative recall tests, and their use of case is impervious to factors influencing message uncertainty or ambiguity, observations which are inconsistent with their hypothesis. However, we find good evidence that participants' behaviour in actual communicative interaction *is* 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,

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

1949; Comrie, 1989; Jäger, 2007). 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, suggesting that Differential Case Marking is the product of a trade-off between economy of effort and communicative considerations penalising the ambiguity or uncertainty about the speaker's message experienced by a listener (e.g. Comrie, 1989; Kurumada & Jaeger, 2015), since marking is restricted to events that are potentially ambiguous or where the speaker's message is assigned low prior probability.

This approach to explaining Differential Case Marking therefore resides in language use, maximising communication (i.e. minimising uncertainty or 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 an artificial language learning experiment 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 result, and 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 communicative considerations 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
'The car avoided the cyclist'

(2) German (Indo-European)

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

[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 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-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

(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'

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 or at least reduce a listener's uncertainty; since

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.

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 communication and economy (Jäger, 2007), marking only when the risk of communication breakdown is high (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 listener uncertainty or 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, 2021, 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 communication-based 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 (2021) 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. Furthermore, Kurumada & Jaeger (2015) show that Japanese speakers are slightly more likely to produce (optional) object case markers in sentences which translate to "The police officer attacked the criminal" than "The criminal attacked the police officer", when the subject is already case-marked as the subject, interpreting this as evidence for a communicative efficiency account where speakers prefer to use additional marking on sentences where the intended message might be relatively improbable, even if intended participant roles are relatively unambiguous. 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, 2023, for review): for instance, this approach has been adopted for morpheme order (St. Clair et al., 2009; Saldana et al., 2021), 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 learned 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 uncertainty (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 uncertainty 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 increased uncertainty or potential 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): even aspects of language design which look like adaptations for communication are in fact a reflection of biases in learning. If so, this would offer an avenue to simplify theories of recurring features of language design (language design would then reflect biases in learning alone, rather than a compromise between partially-conflicting biases in learning and use), while posing important new questions regarding the source of those biases in learning.

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. The first four of these concerns arose from our reading of FNJ, the fifth we only became aware of after running our experiment and receiving feedback from Masha Fedzechkina on an earlier draft of this paper.

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 produce 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 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 associated with interpretational uncertainty or ambiguity. 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 associated with genuine uncertainty as to their interpretation 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 reliably select the intended interpretation in the absence of case markers. On a sceptical reading, 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 crucially depends on uncertainty or ambiguity in its absence. On a less sceptical reading, we might at least expect to see stronger evidence for Differential Object Marking if unmarked animate objects were associated with greater uncertainty as to the intended interpretation.

Fourth, FNJ's Experiment 1 target language includes several statistical features which might plausibly have affected participants tendency to differentially mark animate objects, since they potentially interact with an independently-attested tendency for participants to condition their word order choices on animacy. This concern hinges on the widely-observed tendency to produce animate or human entities before inanimate or non-human entities (e.g. Meir et al., 2017). In the context of FNJ Experiment 1, this should favour SOV order with inanimate objects (animate subject mentioned before inanimate object) and a more even balance between SOV and OSV order with animate objects (both subject and object are animate, so either can be mentioned first). We saw exactky this tendency to use more SOV order with inanimate objects in our pilot experiment based on FNJ Experiment 1. Elsewhere, Fedzechkina et al. (2017) (in an experiment featuring only animate objects, therefore not speaking to Differential Object Marking) report that participants tend to prefer to casemark objects in sentences with OSV order. 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 and is in the minority in the input participants receive) is marked. This preference to use casemarking in OSV order is potentially problematic when it comes to identifying the cause of any preference to mark animate objects, since some of the Differential Object Marking-like effects reported in FNJ may be due to this tendency to use OSV order more with animate objects and mark OSV order.[7] Tal et al. (2022) provide a demonstration of how word order can mediate the occurrence of Differential Object Marking in this way:

[7]While FNJ include word order as a predictor in their statistical model and report an independent effect of animacy on casemarking, since animacy may influence participants' word order choice, these predictors are likely to be somewhat colinear, making estimating their independent effects difficult.

in their experiment, based closely on the method we report here but featuring only animate objects, given (i.e. previously-mentioned) objects are more likely to be mentioned first, and since participants prefer to casemark object-initial sentences, this leads to preferential marking of given objects; since given objects are more likely to be animate, this potentially triggers the emergence of Differential Object Marking. Further complicating this fourth concern, FNJ's participants saw different frequencies of casemarking for the two word orders in their input: casemarking occurred on two thirds of SOV sentences, but on only half of OSV sentences. This may suppress Differential Object Marking, if inanimate objects are more likely to be produced in the SOV order associated with more casemarking in the input. The net effect of these interactions between the human-first bias, the preference to mark OSV order, and conditioning of casemarking on word order in the input is unclear, but make it desirable to at least explore a wider range of input language configurations, including languages where OSV is not the minority order and/or not associated with less frequent casemarking.

Finally, although not reported in the published paper, in FNJ's experiments the experimenter was present in the room with the participant as they learned the artificial language (Fedzechkina, personal communication): the experimenter sat behind the participant and provided feedback during noun production trials, but did not interact with participants on other trials. It could be that the experimenter's presence effectively made the putatively non-communicative recall test in FNJ a communicative test, where the experimenter is a potential audience for the participant's productions, in which case their results would simply be evidence of a bias for efficient communication in a particular (slightly odd) communicative scenario, and do not provide evidence for the efficient-communication-in-learning interpretation they present. If so, the experimental evidence that we provide here, comparing a more straightforward learning-and-recall task with an explicitly communicative task, becomes particularly valuable in locating the bias in learning versus communicative use.

1.5. Our experiment and predictions

We ran artificial language learning experiment targeting Differential Object Marking (henceforth DOM), modelled on FNJ Experiment 1, where we set out to address these issues. We made three changes to the design. First, we train participants on a wider range of input language configurations, including inout languages where OSV is not the minority order and/or not associated with less frequent casemarking, potentially providing a clearer signal of the independent effects of any learning bias for DOM.[8] Second, we manipulate event structure: for some participants unmarked animate objects will pose genuine uncer-

[8]We did not run all 4 input language conditions simultaneously. We initially ran a separate N=47 pilot experiment with majority SOV order and 60% casemarking. This failed to replicate FNJ's DOM result. We then replicated that experiment with a separate and larger sample, which constitutes part of our data reported here (the SOV-majority, 60% casemarking language; note that this produces similar results to our original pilot). We than ran a version with SOV majority word order and 40% casemarking, again failing to find a DOM effect, before running participants on languages with OSV order and 40% or 60% casemarking in order to fully explore the order x casemarking possibilities. Our decision to run the subsequent language types was therefore contingent on the absence of DOM in the first two input languages we tested; however, the absence of those effects is quite consistent across input languages and it is not the case that, e.g, the languages we ran later show clear DOM-like effects which are obscured by our early data showing no effects.

tainty about the intended interpretation. Finally, we add a communicative test: as well as simply producing descriptions for a series of events in a non-communicative recall test, as in FNJ's method, on day 4 participants use the language to communicate with a simulated interlocutor. Our experiment produces a rather different pattern of results to FNJ Experiment 1, and very little evidence supporting a bias for DOM in learning: we see no evidence of a consistent tendency to preferentially mark animate objects, and no influence of the presence or absence of the ambiguity of unmarked animate objects. However, participants rapidly switch to a DOM configuration in communicative interaction, where the uncertainty or ambiguity encountered by the listener becomes relevant. 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.

2. Method

2.1. 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: 522; Day 2: 404; Day 3: 375; Day 4: 341.[9]

2.2. 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. Our target languages were closely based on those from FNJ Experiment 1, but we manipulated majority word order and proportion of casemarking (and therefore the conditioning of casemarking on word order) between participants. 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; case was optionally marked on objects (by the addition of a suffix). In the target language used by FNJ, SOV occurred more than OSV order (60% SOV, 40% OSV), and objects were casemarked more frequently than not (60% of objects were casemarked). We systematically varied both majority word order (in SOV-majority languages: 60% SOV, 40% OSV; in OSV-majority languages: 40% SOV, 60% OSV)

[9]Note that the number of participants declines over days as not all participants return for all 4 days. Of the participants who chose to return, some were blocked from continued participation due to failing noun or sentence comprehension tests: 12 failed the noun comprehension test (see below) on day 1; 25 failed the sentence comprehension test (see below) on day 1; 2 failed the sentence comprehension test on day 2; 2 failed the sentence comprehension test on day 3. We excluded any participants who did not complete all 4 days of the experiment from all analyses, which reduced convergence issues with statistical models where we needed to estimate by-participant random slopes for the effect of day.

Table 1: The 4 target languages. Annotations give the proportion of the sentences in the target language exhibiting each sentence form. There are four possible sentence forms, exhibiting two word orders (SOV and OSV) and the presence or absence of a suffix (-ka) on the object. The majority word order occurs in 60% of sentences. As per FNJ, casemarking and word order are correlated, but this correlation differs across the 4 languages. In the SOV majority languages, SOV sentences are more likely to be casemarked (in the 60% casemarking language) or less likely to be casemarked (in the 40% casemarking language) than OSV sentences. In the OSV majority languages, OSV sentences are more likely to be casemarked (in the 60% casemarking language) or less likely to be casemarked (in the 40% casemarking language) than SOV sentences.

	SOV majority		OSV majority	
Sentence Form	60% case	40% case	60% case	40% case
$N_{subject}$ N_{object}-*ka* V	0.4	0.2	0.2	0.2
$N_{subject}$ N_{object} V	0.2	0.4	0.2	0.2
N_{object}-*ka* $N_{subject}$ V	0.2	0.2	0.4	0.2
N_{object} $N_{subject}$ V	0.2	0.2	0.2	0.4

and frequency of case-marking (60% of objects casemarked or 40% of objects casemarked), producing 4 possible target languages (SOV-majority, 60% case; SOV-majority, 40% case; OSV-majority, 60% case; OSV-majority, 40% case). The grammars of the 4 target languages are given in Table 1.

Rather than using animations as in FNJ, we used drawings to depict events.[10] These drawings included a set of 15 referents (10 animates: artist, boxer, burglar, chef, clown, 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. The target language consists of up to 15 nouns (*slagum, tombat, nagid, melnog, norg, daf, plid, klamen, dacin, zub, vams, bliffen, rungmat, lombur, groost*) and 4 verbs (*slergin, prog, shen, zamper*); 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).

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.

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.

[10]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.

2.3. Manipulation of event composition

As discussed above, in FNJ Experiment 1 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).

2.4. 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 1(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 Subjects 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 1(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 1(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 1(f)); Figure 1(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.[11] 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,[12] and feedback

[11]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).

[12]On each trial where Smeeble was matcher, the software generated a matcher array composed in the

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 subjects 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 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.

2.5. 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 30% of sentence test trials on day 2, falling to 15% 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.[13] 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.

---

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.

[13]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.

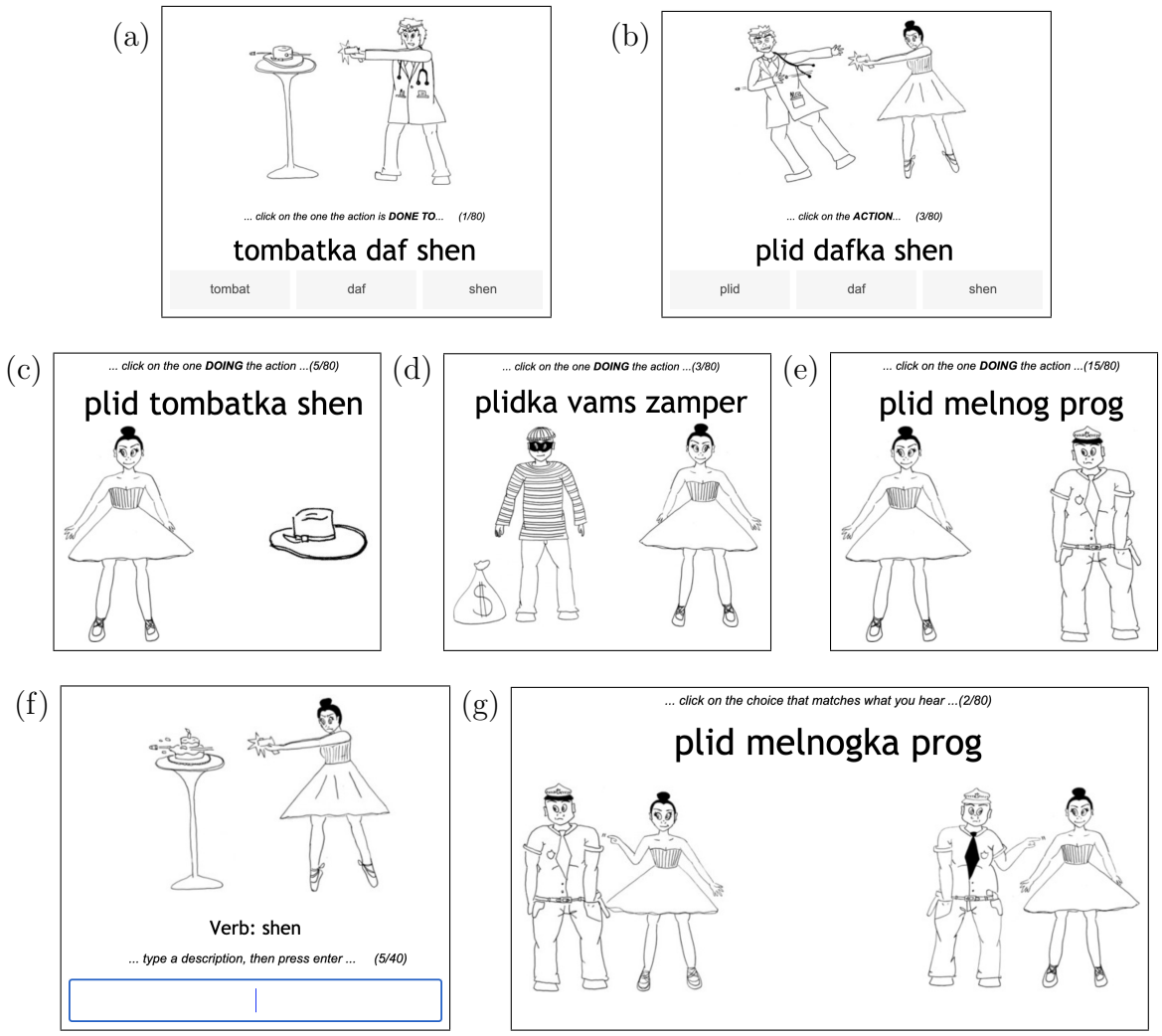

Figure 1: Example sentence training trials (a-b), sentence comprehension trials (c-e), a sentence production trial (f; this trial type also occurs during interaction) and a matcher trial during interaction (g). In training trials participants are required to select one word from the multi-word sentence. 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.

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

3. 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 interpretability of unmarked objects? If so, and if DOM is driven by communicative concerns, then we might expect stronger DOM effects in the Subjects Can Be Objects condition, where we expect unmarked animate objects to be associated with greater uncertainty of interpretation.

We then explore the impact of animacy on word order choices (are participants more likely to produce OSV sentences when describing events involving an animate object?) and whether participants successfully learned the conditioning of case on word order present in their input — as discussed above, these factors and their interaction plausibly affect participants' tendency to casemarked animate objects, potentially affecting our interpretation of those results.

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).

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, 2023) using logistic mixed effects regression using the lme4 package version 1.1-21 (Bates et al., 2015).

3.1. 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 reliably 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 2 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, many participants answer *all* such trials correctly, indicating that the structure of the event set renders unmarked animate objects straightforwardly interpretable and 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 (1-4), casemarking (unmarked or marked), event composition (Subjects Cannot Be Objects or Subjects Can Be Objects) and their interactions (see Table 2 in Appendix A for details of contrasts coding, random effects, and full summary table, and Appendix B for figures and statistics for a full model including frequency of input majority word order and input case marking). 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 2.1 to 1 (log odds of 0.73, $SE = 0.05$), 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 also higher than could be achieved if participants made optimal use of word order cues to agency (input languages use their majority word order, SOV or OSV, 60% of the time, which would yield an expected log odds of success for an attentive learner of 0.41, which is significantly lower than the success rates of our participants; $p < .001$). 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: $b = 0.97, SE = 0.08$; day=3: $b = 2.04, SE = 0.12$; day=4: $b = 2.63, SE = 0.15$). Our data therefore provide strong evidence that participants in the Subjects Cannot Be Objects condition were aware from day 1 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, $b = 1.40$, $SE = 0.13$, $p < .001$, and no significant negative interactions at subsequent days) and particularly marked on days 2 and 3 (as indicated by positive casemarking x day interactions; day 2: $b = 1.24$, $SE = 0.20$, $p < .001$; day 3: $b = 0.52$, $SE = 0.23$, $p = .024$); the absence of a significant positive interaction for day 4 presumably reflects the fact that performance by that point 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 ($b = -0.67$, $SE = 0.07$, $p < .001$), with a log odds of correct responses of close to 0, consistent with chance performance;

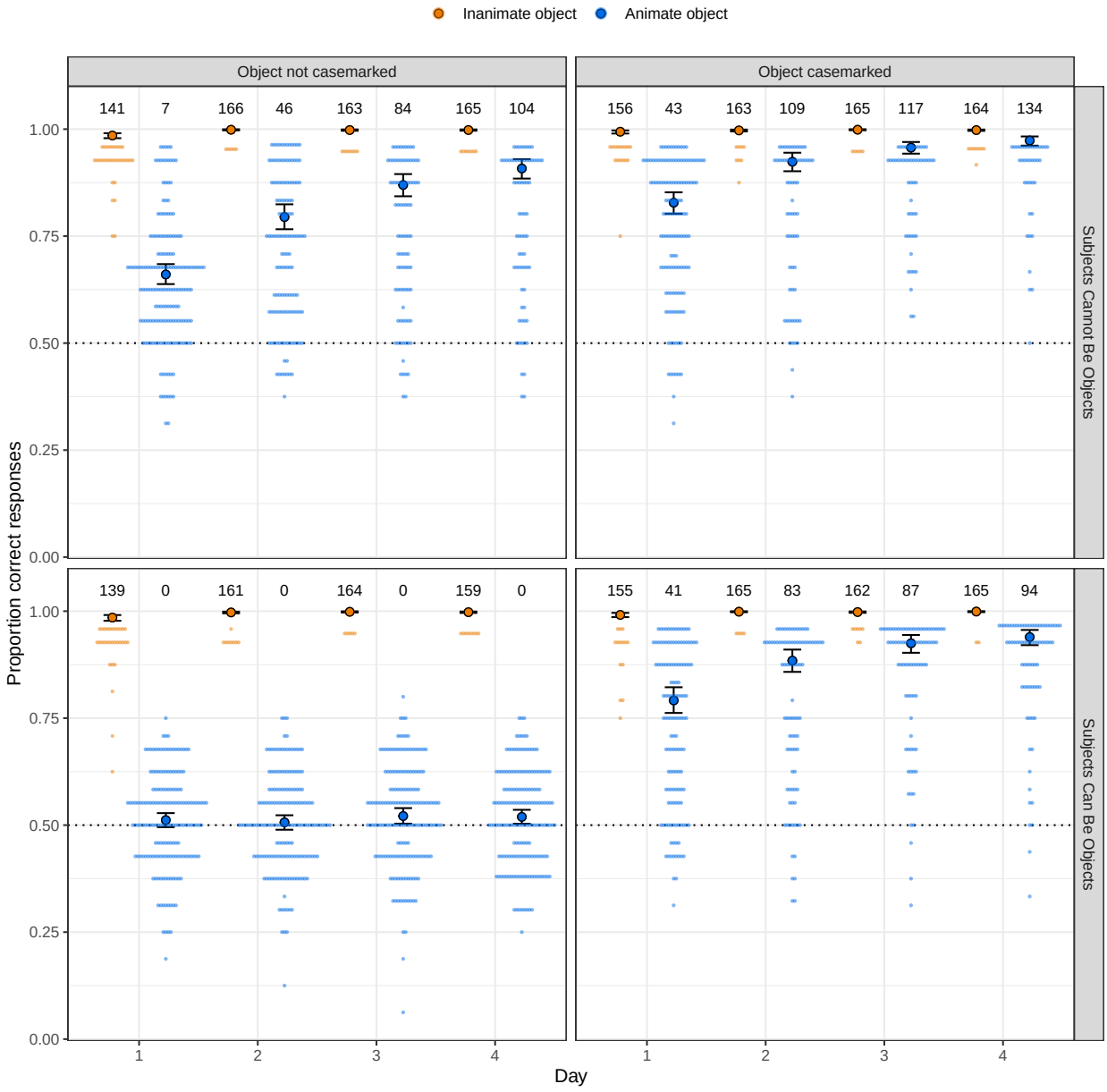

Figure 2: Proportion of correct responses on sentence comprehension trials. Dots give individual participant means. There are too many participants performing at ceiling (i.e. all correct responses) to plot; instead, the text annotations give the number of participants producing ceiling performance. Larger points give mean of by-participant means, error bars give bootstrapped 95% CIs on those means, dashed line indicates chance 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.

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, the Subjects Can Be Objects condition shows a significantly higher benefit from casemarking on day 1 than the Subjects Cannot Be Objects condition (significant casemarking x event composition interaction: $b = 0.36$, $SE = 0.18$, $p = .042$), and this difference becomes larger at days 3 and 4 as the benefit of casemarking declines in the Subjects Cannot Be Objects condition declines (as indicated by significant 3-way interactions involving day 3 and day 4; interaction at day=2: $b = 0.26$, $SE = 0.22$, $p = .239$; interaction at day=3: $b = 1.16$, $SE = 0.26$, $p < .001$; interaction day=4: $b = 1.51$, $SE = 0.32$, $p < .001$).

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

3.2. Animacy affects word order

Recall that we expected that animacy of objects would affect participants' choice of word order, based on the well-documented tendency to mention animates (particularly humans) before inanimates. We saw this effect, although it depended on the majority word order participants encountered in their input. Figure 3 shows how often participants produced SOV word order during the sentence production test phase of the experiment, across days 2–4 of the experiment, broken down by animacy of the object in the event they were describing. Participants trained on SOV-majority input languages are less likely to use SOV order (and therefore more likely to use OSV order) when the object is animate, particularly on Day 2, but participants trained on OSV-majority languages seem to show no such preference.

We ran a logit regression on this word order data, predicting the probability of SOV order using a model with fixed effects of day (2–4), animacy of object (inanimate or animate), input majority word order (OSV or SOV) and their interaction; see Table 3). Participants in the OSV-majority input languages, where SOV sentences form 40% of their input, have a preference for SOV order on day 2 (as indicated by a significant model intercept, indicating participants are producing SOV order at greater than 50%: $b = 0.68$, $SE = 0.24$, $p = .004$); this preference is reduced at days 3 and 4 (as indicated by significant negative effects of day; day 3: $b = -0.69$, $SE = 0.18$, $p < .001$; day 4: $b = -0.64$, $SE = 0.20$, $p = .001$). These participants have no preference to avoid SOV order with animate objects (n. s. effect of animacy: $b = -0.18$, $SE = 0.15$, $p = .218$). Participants trained in SOV-majority languages use SOV order more as expected ($b = 1.95$, $SE = 0.33$, $p < .001$), but also show an effect of animacy on word order, using SOV order less (i.e. OSV order more) when the object is animate (as indicated by a significant negative interaction between SOV-majority inpout and object animacy, $b = -0.83$, $SE = 0.21$, $p < .001$).

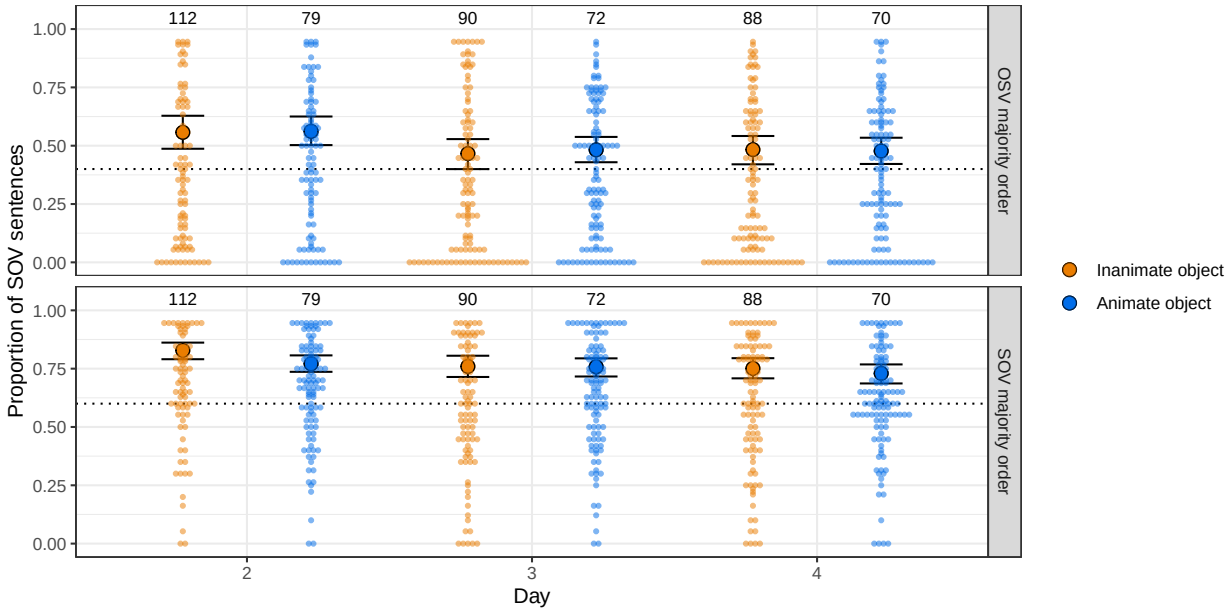

Figure 3: Proportion of SOV (as opposed to OSV) word order in participants' productions (collapsing across event compositions and proportion of casemarking in the input). There are too many participants at 100% SOV to plot, so those participants are indicated in the text annotations. The horizontal dotted line shows frequency of SOV in the input (for both animate and inanimate nouns). Participants over-produce SOV order across all 4 days, and participants trained on SOV-majority languages 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).

3.3. Conditioning of casemarking on word order

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, broken down by the word order participants chose to use. Strikingly, participants appear not to be able to reproduce the conditioning of casemarking on word order present in their input for the two input languages where SOV sentences are casemarked more frequently than OSV sentences (OSV-majority word order and 40% casemarking; SOV-majority order and 60% casemarking), but can reproduce this conditioning when trained on input languages which mark OSV order more than SOV order (OSV-majority word order and 60% casemarking; SOV-majority order and 40% casemarking).

Table 4 gives the accompanying analysis from a logistic regression predicting presence of a casemarker based on fixed effects of day (2-4), word order (OSV or SOV), proportion of casemarking in the input (40% or 60%) and which word order is marked more frequently in the input (SOV or OSV)[14] and their interactions. For input languages where SOV order in marked more there is no effect of word order ($b = -0.08$, $SE = 0.22$, $p = .708$), but for input languages where OSV is marked more than SOV there is (as indicated by a significant interaction between word order and which order is marked more: $b = -1.79$, $SE = 0.32$, $p < .001$).

This result is consistent with the observation by Fedzechkina et al. (2017) that participants preferentially use casemarking with OSV order — our participants succesfulyl reprdocued the conditioning of case on word order when it favoured marking OSV order, but not when it favoured marking SOV order. As noted above, since participants are more likely (at least for SOV-majority languages) to use OSV order with animate objects, 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 OSV); designing paradigms which cleanly separate all these factors is challenging. This also suggests that the clearest signature for efficient communication biases might be found in the OSV-majority language with 40% casemarking, since those participants' use of casemarking is approximately equal for SOV and OSV sentences, and those participants do not show the tendency seen elsewhere to preferentially produce OSV sentences with animate objects.

3.4. Animacy and casemarking during sentence recall: no evidence of a DOM-like effect

Figure 5 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. The results do not look like FNJ Experiment 1: participants' production of case marking is, on average, consistent with the frequency of casemarking in their input, even on day 2, and there is no consistent tendency to casemark animate objects more on any day. There is no obvious effect of event composition on casemarking, despite its very clear and strengthening effect on the interpretabilty of unmarked animate objects.

[14]Using this predictor rather than majority word order simplifies the interpretation of the results, since otherwise the difference seen in Figure 4 depends on the interaction between proportion of marking and majority word order.

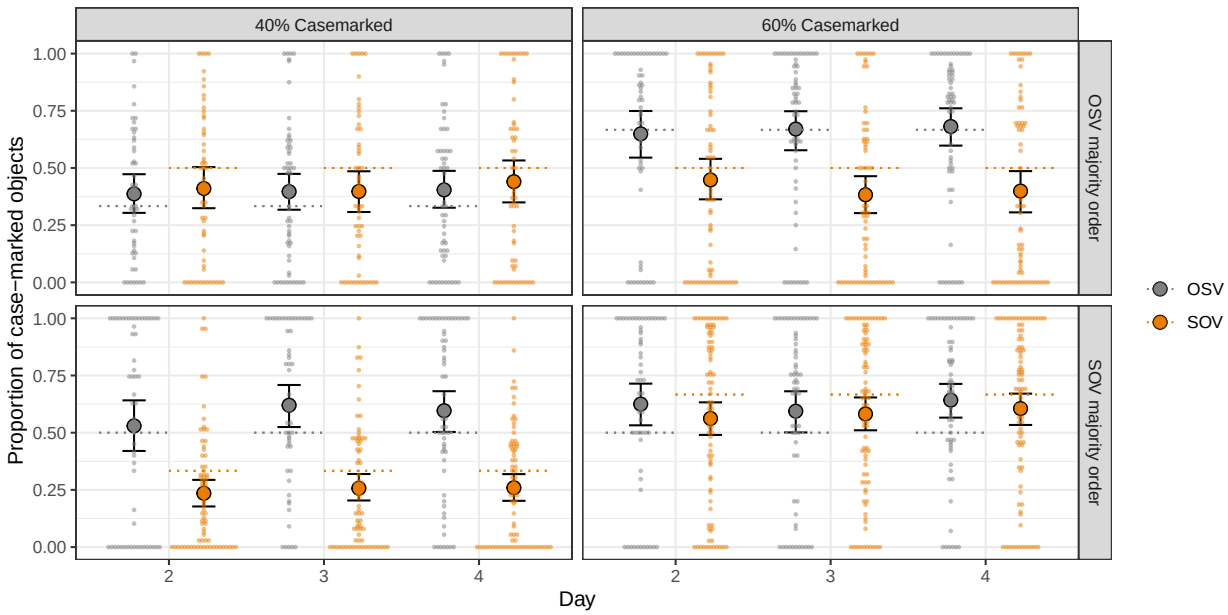

Figure 4: Proportion of casemarking in participants' productions, broken down by word order. The horizontal dotted line shows case frequency of casemarking in the input, which differs between the two orders in all 4 input language configurations. When the input features more frequent casemarking of OSV sentences (in the OSV-majority language with 60% casemarking, or the SOV-majority language with 40% casemarking) participants reproduce this conditioning of casemarking on word order; however, when the input marks SOV sentences more, participants fail to reproduce that conditioning.

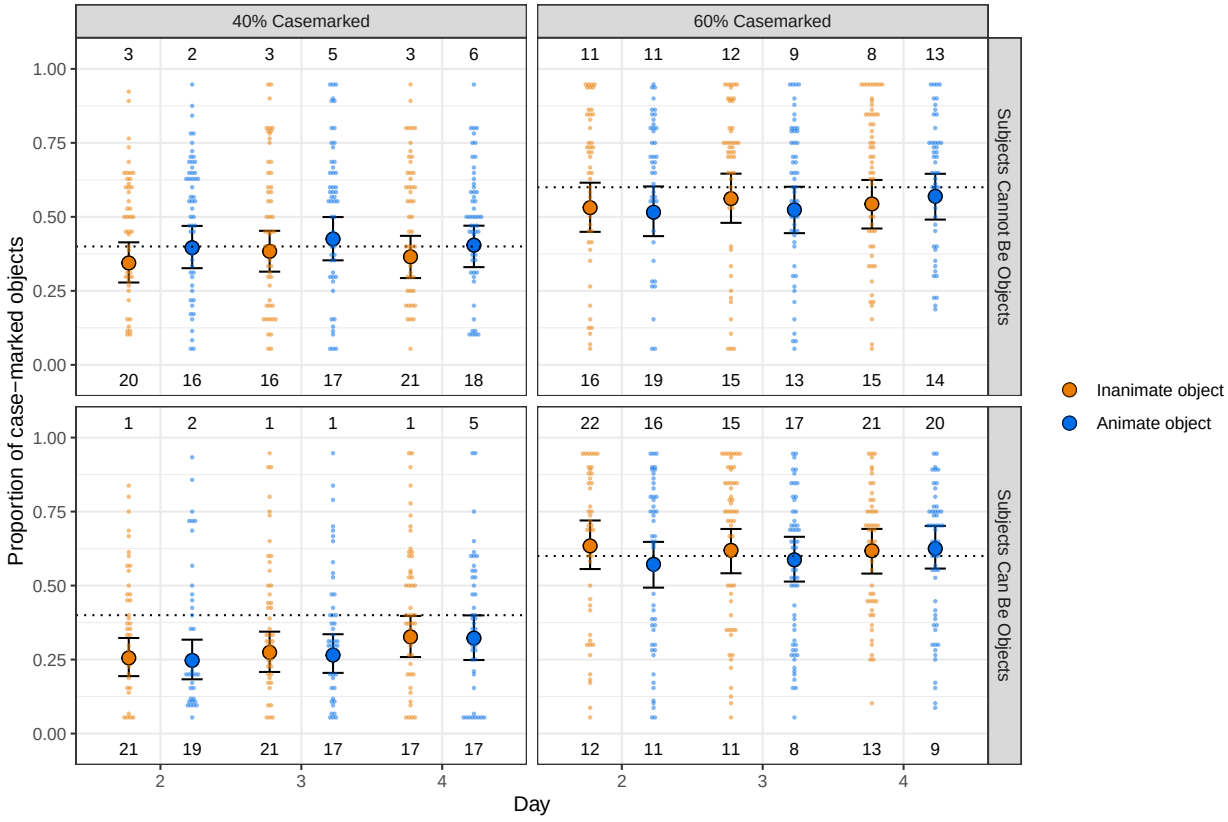

Figure 5: Proportion of casemarking in participants' productions. The horizontal dotted lines shows casemarking frequency in the input (for both animate and inanimate nouns). Contrary to the predictions of the efficient-communication-in-learning hypothesis, there is no consistent relationship between animacy and casemarking, and no reliable tendency to over-mark animate objects.

We analysed this data using a logistic regression, predicting presence of a casemarker based on fixed effects of animacy (inanimate or animate object), event type (Subjects Cannot Be Objects, Subjects Can Be Objects), day (2-4), proportion of casemarking in the input (40% or 60%), which word order is marked more frequently in the input (SOV or OSV)[15] and their interactions. See Table 5 for details of contrast coding and a full summary table of fixed effects.

There is no consistent DOM-like bias in production at day 2 (as indicated by an n.s. effect of animacy, $b = -0.01$, $SE = 0.12$, $p = .950$) and no interaction between animacy and day=4 which would indicate a shift towards a more DOM-like configuration over 4 days ($b = 0.09$, $SE = 0.13$, $p = .478$). Furthermore, there are no interactions between animacy and event composition, which we think would be a reasonable prediction under the efficient communication account (this would predict a positive interaction, e.g., more marking of animate objects in the Subjects Can Be Objects condition at day 2; instead we see a negative but n.s. interaction, $b = -0.33$, $SE = 0.23$, $p = .159$). 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.

There is however an interaction between animacy and proportion of casemarked objects in the input language ($b = -0.51$, $SE = 0.24$, $p = .030$) which suggests, 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*. We return to possible explanations for this unexpected finding in the discussion. This effect is however eliminated by day 4 (as indicated by a significant three-way interaction between animacy, proportion of casemarking in the input, and day=4: $b = 0.53$, $SE = 0.26$, $p = .036$).[16]

We conducted several additional analyses to verify the absence of DOM-like effects in places of particular interest. Re-running the same main analysis with Day 4 set to the model intercept allows us to verify the absence of a DOM-like effect of animacy on day 4 ($b = 0.08$, $SE = 0.11$, $p = 0.461$). The SOV-majority, 60% casemarking language closely matches the target language used in FNJ Experiment 1; when analysing just this data with a model featuring day, animacy and event composition we again see no DOM-like effect of animacy on day 4 ($b = 0.34$, $SE = 0.22$, $p = 0.113$) and no interaction between animacy

[15] We include this as a predictor rather than majority order since, as shown elsewhere, it modulates participants' tendency to condition casemarking on word order

[16] The only other significant effects are the expected effect of proportion of casemarking in the input (with more casemarking in the input leading to more casemarking in recall: $b = 1.97$, $SE = 0.33$, $p < .001$), an increase in casemarking on day 4 ($b = 0.29$, $SE = 0.13$, $p = .028$; the corresponding effect on day 3 is marginal, suggesting a progressive increase in use of casemarking across the 3 testing days) and two unexpected interactions involving event composition: we see more casemarking in the Subjects Can Be Objects condition when casemarking is frequent in the input, and less casemarking in the Subjects Can Be Objects condition when SOV order is casemarked more than OSV in the input. Since neither of these interactions involve animacy they do not relate to Differential Object Marking; since they depend on statistical properties of the input language they do not have an obvious communication-based explanation either (e.g. it is not simply the case that we see more casemarking of objects in the Subjects Can Be Objects condition) and we do not attempt to interpret them further.

and event composition. As discussed above, the OSV-majority, 40% casemarking language provides another opportunity to test for DOM-like effects, since participants in this condition should show less tendency to condition word order on animacy of the object or condition casemarking on word order; again, we see no DOM-like effect of animacy on day 4 ($b = -0.20$, $SE = 0.25$, $p = .428$) and no interaction between animacy and event composition.

3.5. Differential Object Marking is present in communicative interaction

On day 4 of the experiment, after completing their final sentence recall test, participants use the language they have learned to play a communication game with Smeeble, their monster language tutor.[17] Figure 6 shows day 4 data, contrasting the (non-communicative) pre-interaction recall test with behaviour during interaction; Table 6 gives the accompanying stats, where we use a logistic regression to predict participants' use of casemarking based on fixed effects of block (recall or interaction), animacy of object, event composition, proportion of casemarked objects in the input, and their interaction. This analysis indicates no effect of animacy during pre-interaction sentence recall ($b = 0.16$, $SE = 0.11$, $p = .168$), 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 ($b = 0.65$, $SE = 0.11$, $p < .001$), indicating that participants casemarked more during interaction, and a significant interaction between block and animacy ($b = 0.67$, $SE = 0.11$, $p < .001$), indicating that this increase in marking was greater for animate objects, i.e. while there is very little evidence that a DOM-like configuration emerges from learning, it rapidly develops during communication when the communicative function of casemarking becomes relevant. There is also a significant 3-way interaction suggesting that this increase in differential marking of animates is smaller in the 60% casemarking languages ($b = -0.55$, $SE = 0.22$, $p = .014$), possibly due to more participants already being at ceiling case marking in pre-interaction recall; however, a relevelled model taking the 60% casemarked languages as the reference level indicates that participants trained on these languages still show a significant block x animacy interaction indicating a DOM-like behaviour in the interaction block ($b = 0.40$, $SE = 0.15$, $p = .008$).

This analysis also reveals a significant increase in differential casemarking in the Subjects Can Be Objects condition in interaction (as indicated by a significant three-way interaction between block, animacy and event composition: $b = 0.57$, $SE = 0.21$, $p = .007$), suggesting that participants increase their use of differential marking in the condition where, in their experience, unmarked animates pose genuine interpretability challenges. 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 running our experiment that participants would find unmarked

[17] We omit the graphs and analysis for brevity, but as in the analysis of sentence comprehension trials, during the interaction phase 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.

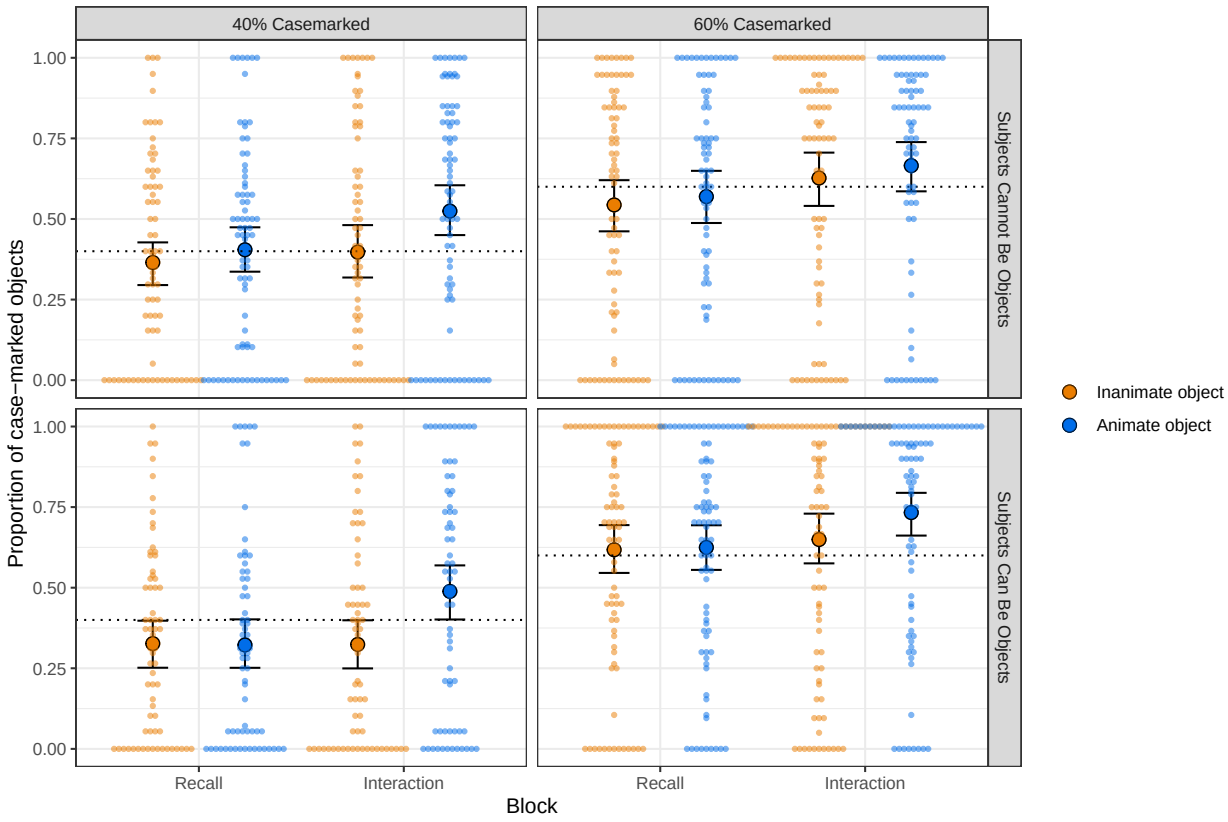

Figure 6: Frequency of casemarking in participants' productions on day 4, during the sentence production test (labelled 'Recall' here) and subsequent interaction with Smeeble. The horizontal dotted line shows casemarking frequency in the input (for both animate and inanimate nouns). While there is no 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 communicative challenges for participants.

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 interpretability 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 communicative requirements of their interlocutor.

4. Discussion

4.1. DOM in communication, but not in learning

We see essentially no clear evidence supporting the efficient-communication-in-learning account proposed by FNJ: participants do not reliably condition their use of casemarkers on the animacy of the object in the non-communicative sentence production trials, and are impervious to the in-practice interpretability of unmarked animates (which differs based on our manipulation of event composition).[18] We do however see very clear effects of communicative use of the language on casemarking: participants switch to a DOM configuration where animate objects are preferentially marked. 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.

4.2. Methodological differences from FNJ

While our method was based on that reported by FNJ, there are some methodological differences arising from conducting data collection online that may be consequential.[19] First, as discussed above, although not reported in the published paper, in FNJ's experiments the the experimenter was present in the room with the participant as they learned the artificial language and potentially acted as an audience to whom the participant addressed their utterances. This potentially which undermines what we regard as the critical claim in their paper, that biases for efficient communication operate *in learning*.

In FNJ, training was passive: participants watched short videos accompanied by their auditory descriptions. In our study training was active: on every trial, participants provided a response, e.g. in sentence training they were required to click on the word referring to the the action, the doer of the action, or who the action was done to, with nouns presented without case marking. We also instructed participants to pay attention to the words and the order they were in, in response to heavy over-use of SOV word order and SVO order in an early pilot experiment. This combination of design decisions might have led participants to pay more attention to sentence word order over casemarking, although it is worth noting

[18]Note that we see quite clear effects of animacy on word order during the learning phase of the experiment, and participants distinguish between animates and inanimates during actual communicative interaction. This 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; 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.

[19]We are grateful to Masha Fedzechkina for enumerating these differences and their potential consequences in response to an earlier draft of this manuscript.

that participants already seem to understand the function of the case marker in day 1 (as indicated by the effect of casemarking on accuracy in comprehension), and have no problem in using it to disambiguate where required in the communicative interaction phase on day 4; in contrast they show only limited evidence of using word order to attempt to disambiguate unmarked animates, and only when casemarking is less frequent. The instruction to attend to word order (and/or the smaller lexicon we used relative to FNJ) might also have helped participants learn that subjects and objects were non-overlapping sets in the Subjects Cannot be Objects condition, whereas FNJ's participants apparently failed to learn this.

4.3. A bias for typicality matching in learning?

The one place where we see fleeting evidence for any effect of animacy on casemarking is on day 2 of the experiment, in participants' first sentence production test; this effect is modulated by the frequency of casemarking in the participants' input, such that participants trained on languages where the minority of objects are casemarked show a slight preference to mark animate objects (a DOM-like effect), whereas participants trained on languages where the majority of objects are casemarked show a slight preference to mark inanimate objects (an anti-DOM configuration). This effect is extinguished after 4 days of training, and could simply reflect a chance fluctuation in our (admittedly rather messy) data. However, another 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 whether casemarked object nouns are more frequent than unmarked object nouns or not. If participants initially assumed that the more typical object type (inanimate) should be marked with the more frequent/typical object noun form, this would yield a DOM-like effect in minority-casemarked languages, and an anti-DOM effect in majority-casemarked languages; if participants were subsequently able to learn, as is the case in their input, that object typicality does not align with noun form frequency then this would predict the loss of this alignment, 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.

4.4. 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 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 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*-∅ 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*-∅, 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.[20]

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 communicative 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[21], 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

[20]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.

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

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.[22]

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 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).[23]

[22]We understand that the presence of the experimenter in the room with the participant was Fedzechkina's standard practice in lab-based experiments (Fedzechkina, personal communication), which means the same critique outlined above regarding whether their results speak to biases in learning or communication applies to Fedzechkina et al. (2017), run in the lab, but not Fedzechkina & Jaeger (2020), which shows the same phenomenon but is run online.

[23]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

4.5. 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, 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 analysis of the communication phase of our experiment, we found some evidence that participants' behaviour during interaction was influenced 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 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, 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 but rather by a preference to mark low-probability messages (either for communicative efficiency or iconicity), it would be interesting to test whether 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

verified this statistically with a model based on our recall and interaction data from day 4: 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 ($b = 0.01, SE = 0.03$). 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.

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.

5. 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. Contrary to Fedzechkina et al. (2012), we find little evidence for a bias in learning favouring communicatively-efficient Differential Object Marking: participants do not reliably condition their use of casemarking on the animacy of objects, and their behaviour is impervious to the communicative challenges arising from unmarked objects. However, we find good evidence that participants' behaviour in actual communicative language use in interaction *is* driven by efficient communication considerations: in interaction participants exhibited the expected Differential Object Marking pattern. 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 our experiments reported here, and Hanna Jarvinen for assisting with stimuli preparation. Thanks to Olga Fehér, Mora Maldonado, and Alan Smith for help with examples in Hungarian, Spanish, and German. Thanks to Masha Fedzechkina for feedback on an earlier draft of this manuscript.

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. (2023). 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.

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. doi:10.1016/j.cognition.2018.12.002.

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. (2021). Role-reference associations and the explanation of argument coding splits. *Linguistics*, *59*, 123–174. doi:10.1515/ling-2020-0252.

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*, *105*, 10681–10686. doi:10.1073/pnas.0707835105.

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.

Kurumada, C., & Jaeger, T. F. (2015). Communicative efficiency in language production: Optional case-marking in Japanese. *Journal of Memory and Language*, *83*, 152–178. doi:10.1016/j.jml.2015.03.003.

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.

R Core Team (2023). *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. (2021). Cross-linguistic patterns of morpheme order reflect cognitive biases: An experimental study of case and number morphology. *Journal of Memory and Language*, *118*, 104204. doi:10.1016/j.jml.2020.104204.

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. doi:10.1016/j.cognition.2010.06.004.

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.

Tal, S., Smith, K., Culbertson, J., Grossman, E., & Arnon, I. (2022). The Impact of Information Structure on the Emergence of Differential Object Marking: An Experimental Study. *Cognitive Science*, *46*. doi:10.1111/cogs.13119.

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.

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 2: Summary table for the statistical analysis of identification accuracy in sentence comprehension trials (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 and random slopes for Casemarking and Day. 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)	0.73	0.05	$< .001^{***}$
Casemarking	1.40	0.13	$< .001^{***}$
Day=2	0.97	0.08	$< .001^{***}$
Day=3	2.04	0.12	$< .001^{***}$
Day=4	2.63	0.15	$< .001^{***}$
Subjects Can Be Objects	-0.67	0.07	$< .001^{***}$
Casemarking * Day=2	1.24	0.20	$< .001^{***}$
Casemarking * Day=3	0.52	0.23	$.026^{*}$
Casemarking * Day=4	0.47	0.28	.096
Subjects Can Be Objects * casemarking	0.36	0.18	$.042^{*}$
Subjects Can Be Objects * Day=2	-1.01	0.10	$< .001^{***}$
Subjects Can Be Objects * Day=3	-2.02	0.15	$< .001^{***}$
Subjects Can Be Objects * Day=4	-2.61	0.18	$< .001^{***}$
Subjects Can Be Objects * casemarking * Day=2	0.36	0.22	.239
Subjects Can Be Objects * casemarking * Day=3	1.16	0.26	$< .001^{***}$
Subjects Can Be Objects * casemarking * Day=4	1.51	0.32	$< .001^{***}$

Table 3: Summary table for the statistical analysis of word order during the sentence production test. We ran a logit regression predicting use of SOV word order based on fixed effects of Day, Animacy, Majority Word Order in the training input, and their interactions, with by-participant random intercepts and by-participant random slopes for Day, Animacy, and their interaction. Day and Majority Word Order are treatment-coded, and Animacy is treatment coded; the model intercept therefore reflects the probability of SOV order at Day 2 for participants trained on the OSV-majority input language, and a positive effect of Animacy indicates greater use of SOV for animate objects.

	b	SE	p
Intercept (Day=2)	0.68	0.24	.004**
Animacy	-0.18	0.15	.218
Day=3	-0.69	0.18	< .001***
Day=4	-0.64	0.20	.001**
Majority Word Order=SOV	1.95	0.33	< .001***
Animacy * Day=3	0.26	0.18	.134
Animacy * Day=4	0.06	0.17	.724
Animacy * Majority Word Order=SOV	-0.83	0.21	< .001***
Majority Word Order=SOV * Day=3	0.48	0.24	.050*
Majority Word Order=SOV * Day=4	0.16	0.27	.554
Animacy * Majority Word Order=SOV * Day=3	0.02	0.26	.940
Animacy * Majority Word Order=SOV * Day=4	0.26	0.25	.285

B. Additional figures and analyses

Figure 7 shows the proportion of sentence comprehension trials on which participants were able to identify the subject, broken down by properties of the input language (majority word order in the input, proportion of casemarked objects), for unmarked animate objects only since these are revealing regarding our manipulation of event composition. All input language configurations show the same key effect as reported in the main text, namely that unmarked animates are not ambiguous in the Subjects Cannot Be Objects Condition even on Day 1, but are ambiguous and remain so across all 4 days in the Subjects Can Be Objects condition. One additional observation from this figure is that participants trained on input languages with less frequent casemarking (40% of objects are marked, rather than 60%) show some evidence of using word order as a cue to subjecthood: this can be seen in most facets in Figure 7, but is particularly clear in the Subjects Can Be Objects conditions, where participants trained on 40% casemarking languages are performing above chance for unmarked objects, suggesting they may be using word order as a probabilistic cue (recall that in all input languages either SOV or OSV order occurs on 60% of sentences).

These impressions are confirmed by a statistical analysis, again on comprehension trials involving unmarked animate objects only, where we predict response accuracy (correct or incorrect) based on fixed effects of day (1-4), event composition (Subjects Cannot Be Objects or Subjects Can Be Objects), majority input word order (OSV or SOV) and proportion of casemarking in the input (40% or 60%) (see Table 7 for details of contrasts coding, random effects, and full summary table). As well as replicating the crucial features of the analysis reported in the main text, this analysis shows an effect of proportion of casemarked objects at Day 1 ($b = -0.36$, $SE = 0.13$, $p = .006$), indicating that participants in the Subjects Cannot

Table 4: Summary table for the statistical analysis of the effect of word order on casemarking. We ran a logit regression predicting use of casemarking based on fixed effects of Day, Word Order, which Word Order was Casemarked More (OSV or SOV) in the input, Proportion of Casemarked Objects in the input, and their interactions with by-participant random intercepts and by-participant random slopes for Day, Word Order, and their interaction. Day and Word Order Casemarked More were treatment-coded, and Word Order and Proportion of Casemarked Objects were deviation coded: the model intercept reflects the probability of casemarking at Day 2 for participants trained in the SOV Casemarked More input language, and positive effects of Word Order / Proportion Casemarked indicate greater use of casemarking for SOV sentences / for participants trained on the 60% casemarked input language. 3- and 4-way interactions are all n.s. (lowest p = .102) and can be found in full in the supporting online information.

	b	SE	p
Intercept (Day=2, Word Order Casemarked More=SOV)	-0.39	0.23	.090
Word Order	-0.08	0.22	.708
Proportion Casemarked	1.65	0.45	$< .001^{***}$
Word Order Casemarked More=OSV	-0.04	0.32	.897
Day=3	0.16	0.15	.278
Day=4	0.34	0.17	.053
Word Order * Day=3	0.11	0.25	.671
Word Order * Day=4	0.11	0.26	.657
Word Order * Proportion Casemarked	-0.41	0.43	.340
Word Order * Word Order Casemarked More=OSV	-1.79	0.32	$< .001^{***}$
Proportion Casemarked * Day=3	-0.10	0.30	.736
Proportion Casemarked * Day=4	-0.06	0.34	.866
Proportion Casemarked * Word Order Casemarked More=OSV	-0.17	0.64	.788
Word Order Casemarked More=OSV * Day=3	-0.08	0.21	.719
Word Order Casemarked More=OSV * Day=4	-0.28	0.25	.253
All higher-order interactions			$> .102$

Table 5: Summary table for the statistical analysis of casemarking 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, Proportion Casemarked objects in the input language (40% or 60%), which Word Order was Casemarked More (OSV or SOV) 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, Event Composition and Proportion Casemarked were deviation-coded; the model intercept therefore reflects the probability of casemarking an object at Day 2 collapsing across Animacy, Event Composition, Proportion Casemarked, and Word Order Casemarked More, and positive effects reflect more casemarking for animate objects / in the Subjects Can Be Objects condition / for participants trained on the 60% casemarked input language / for participants trained on an input language where SOV order is casemarked more. 4- and 5-way interactions are all n.s. (lowest p = .103) and can be found in full in the supporting online information.

	b	SE	p
Intercept (Day=2)	-0.73	0.17	$< .001^{***}$
Animacy	-0.01	0.12	.950
Proportion Casemarked	1.97	0.33	$< .001^{***}$
Event Composition	-0.04	0.33	.893
Word Order Casemarked More	0.24	0.33	.476
Day=3	0.20	0.11	.073
Day=4	0.29	0.13	$.028^{*}$
Animacy * Day=3	-0.01	0.13	.918
Animacy * Day=4	0.09	0.13	.478
Event Composition * Day=3	-0.07	0.22	.760
Event Composition * Day=4	0.22	0.26	.393
Proportion Casemarked * Day=3	-0.17	0.23	.464
Proportion Casemarked * Day=4	-0.08	0.26	.769
Word Order Casemarked More * Day=3	0.12	0.22	.580
Word Order Casemarked More * Day=4	0.29	0.26	.262
Animacy * Event Composition	-0.33	0.23	.159
Animacy * Proportion Casemarked	-0.51	0.24	$.030^{*}$
Animacy * Word Order Casemarked More	-0.14	0.23	.543
Event Composition * Proportion Casemarked	1.63	0.65	$.012^{*}$
Event Composition * Word Order Casemarked More	-1.42	0.65	$.029^{*}$
Proportion Casemarked * Word Order Casemarked More	-0.12	0.65	.859
Animacy * Event Composition * Day=3	0.16	0.25	.531
Animacy * Event Composition * Day=4	0.08	0.24	.742
Animacy * Proportion Casemarked * Day=3	0.12	0.26	.633
Animacy * Proportion Casemarked * Day=4	0.53	0.26	$.036^{*}$
Event Composition * Proportion Casemarked * Day=3	0.01	0.44	.978
Event Composition * Proportion Casemarked * Day=4	-0.48	0.51	.346
Animacy * Event Composition * Proportion Casemarked	-0.10	0.46	.830
Animacy * Event Composition * Proportion Casemarked * Day=3	0.52	0.49	.288
Animacy * Event Composition * Proportion Casemarked * Day=4	0.24	0.48	.612
Animacy * Proportion Casemarked * Word Order Casemarked More	0.32	0.45	.486
Animacy * Event Composition * Word Order Casemarked More	0.21	0.46	.645
Proportion Casemarked * Event Composition * Word Order Casemarked More	-0.39	1.21	.749
Animacy * Word Order Casemarked More * Day=3	0.01	0.25	.955
Animacy * Word Order Casemarked More * Day=4	0.22	0.24	.360
Proportion Casemarked * Word Order Casemarked More * Day=3	0.30	0.44	.495
Proportion Casemarked * Word Order Casemarked More * Day=4	0.23	0.51	.653
Event Composition * Word Order Casemarked More * Day=3	0.51	0.44	.247
Event Composition * Word Order Casemarked More * Day=4	0.12	0.51	.811
All higher-order interactions			$> .103$

Table 6: Summary table for the statistical analysis of casemarking on day 4, 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, Proportion Casemarked in the input language, 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 Casemarked were deviation-coded: the model intercept therefore reflects the probability of casemarking during pre-interaction Recall, collapsing across Animacy, Event Composition, and Proportion Casemarked, and positive effects of Animacy / Event Composition / Proportion Casemarked reflect greater use of casemarking for animate objects / in the Subjects Can Be Objects condition / for participants trained on the 60% casemarked input language.

	b	SE	p
Intercept (Block=Recall)	-0.42	0.16	.010*
Animacy	0.16	0.11	.168
Block = Interaction	0.65	0.11	< .001***
Event Composition	0.19	0.33	.554
Proportion Casemarked	1.86	0.33	< .001***
Animacy * Block=Interaction	0.67	0.11	< .001***
Animacy * Event Composition	-0.20	0.22	.358
Animacy * Proportion Casemarked	-0.17	0.23	.444
Event Composition * Block=Interaction	0.13	0.21	.556
Proportion Casemarked * Block=Interaction	0.32	0.22	.147
Event Composition * Proportion Casemarked	1.04	0.64	.105
Animacy * Event Composition * Block=Interaction	0.57	0.21	.007**
Animacy * Proportion Casemarked * Block=Interaction	-0.55	0.22	.014*
Animacy * Event Composition * Proportion Casemarked	0.07	0.44	.874
Event Composition * Proportion Casemarked * Block=Interaction	0.19	0.43	.658
Animacy * Event Composition * Proportion Casemarked * Block=Interaction	-0.07	0.41	.871

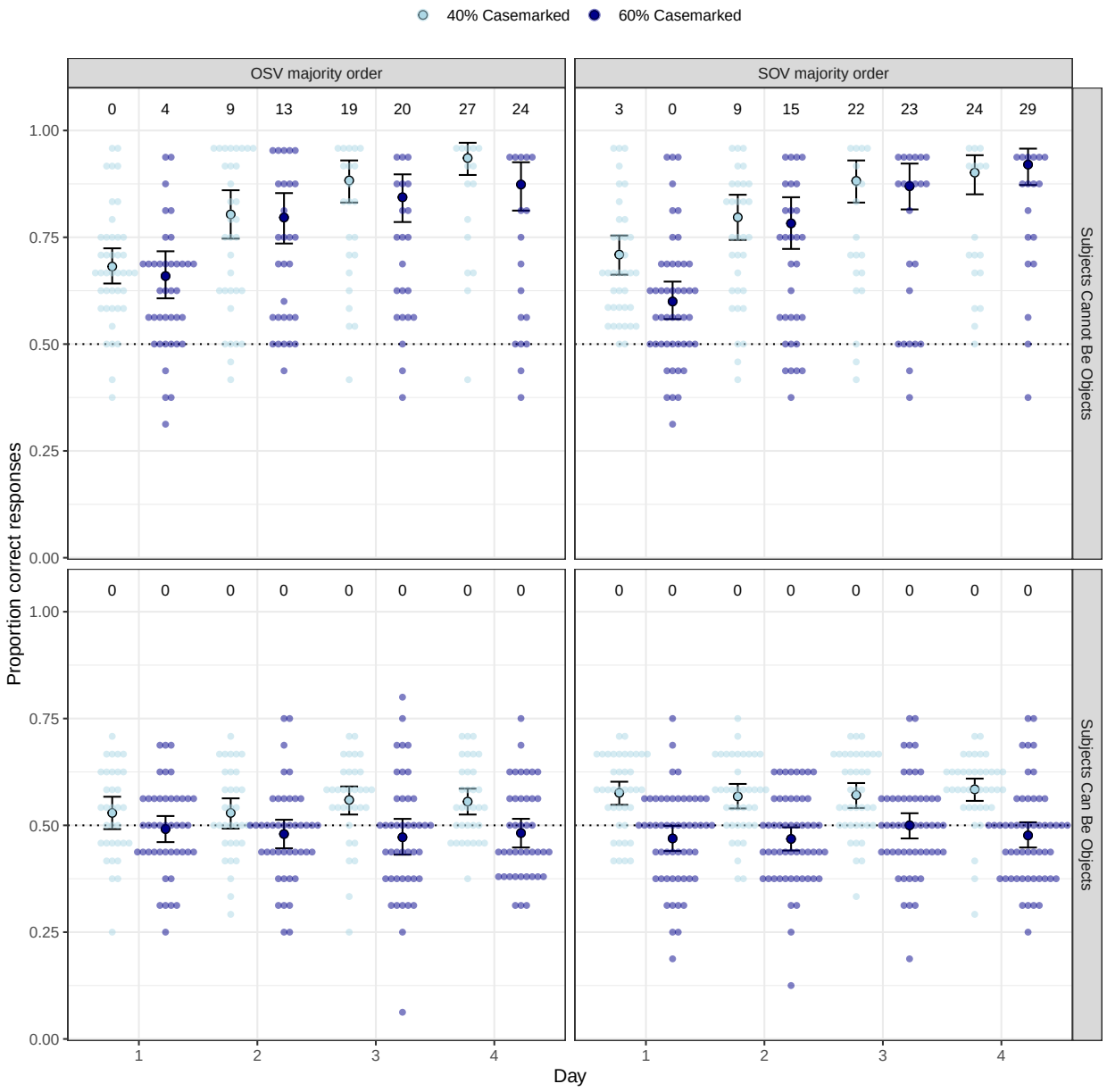

Figure 7: Proportion of correct responses on sentence comprehension trials, for unmarked animate objects only, broken down by properties of the input language (majority word order, percentage of casemarked objects). Plotting conventions as in Figure 2

.

Be Objects trained on 60% casemarked languages are less successful at correctly identifying unmarked animate subjects than participants trained on 40% casemarked languages; the absence of an interaction with event type indicates this effect is present to approximately the same extent in the Subjects Can Be Objects condition. This advantage for less frequent casemarking in the input appears to be reduced at day 2, as indicated by a significant interaction between day=2 and casemarking proportion, but since that interaction is only

present in day 2 and not at days 3 and 4 it is probably not representative of any consistent trend.

There are some more complex interactions involving input language composition. Participants trained on the SOV-majority, 60% casemarked input language in the Subjects Cannot Be Objects condition perform worse than other input language configurations on day 1 (as indicated by a significant interaction between proportion of casemarked objects and majority input order) but this seems to be a quirk of Day 1, as indicated by positive 3-way interactions with Day 2, Day 3 and Day 4 (the Day 3 and Day 4 interactions are significant), and indeed by Day 4 they are the best at correctly identifying unmarked animate subjects. The complex 4-way interaction between day, proportion of casemarking, majority input order and Subjects Can Be Objects simply marks the absence of this effect in the Subjects Can Be Objects condition.

In sum, the only consistent effect of input language properties on sentence comprehension seems to be that participants trained on input languages with less frequent casemarking are more attentive to word order as a potential cue to subjecthood, allowing them to perform better on unmarked animates.

Table 7: Summary table for the statistical analysis of identification accuracy in sentence comprehension trials (unmarked animate objects only). We ran a logit regression predicting accuracy (correct or incorrect) based on fixed effects of Day, Event Composition, Majority Word Order, Proportion Casemarked and their interaction. The model only includes by-participant random intercepts, since including random slopes for Day lead to convergence issues. Day and Event Composition are treatment coded: the model intercept therefore reflects log-odds of correctly identifying the subject on day 1 in the Subjects Cannot Be Objects condition. Majority Word Order and Proportion Casemarked are deviation-coded such that positive effects indicate greater accuracy for SOV majority input order / 60% casemarking in the input. Note: $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

	b	SE	p
Intercept (Day=1, Subjects Cannot Be Objects)	0.81	0.07	$< .001^{***}$
Day=2	0.78	0.06	$< .001^{***}$
Day=3	1.38	0.07	$< .001^{***}$
Day=4	1.83	0.07	$< .001^{***}$
Subjects Can Be Objects	-0.74	0.09	$< .001^{***}$
Proportion Casemarked	-0.36	0.13	$.006^{**}$
Majority Word Order	-0.06	0.13	.645
Subjects Can Be Objects * Day=2	-0.80	0.08	$< .001^{***}$
Subjects Can Be Objects * Day=3	-1.35	0.08	$< .001^{***}$
Subjects Can Be Objects * Day=4	-1.79	0.09	$< .001^{***}$
Proportion Casemarked * Day=2	0.28	0.12	$.023^{*}$
Proportion Casemarked * Day=3	0.07	0.13	.582
Proportion Casemarked * Day=4	0.07	0.15	.631
Majority Word Order * Day=2	0.00	0.12	.988
Majority Word Order * Day=3	0.17	0.13	.203
Majority Word Order * Day=4	0.09	0.15	.532
Subjects Can Be Objects * Majority Word Order	0.11	0.18	.538
Subjects Can Be Objects * Proportion Casemarked	0.06	0.18	.730
Majority Word Order * Proportion Casemarked	-0.54	0.26	$.040^{*}$
Subjects Can Be Objects * Proportion Casemarked * Day=2	-0.28	0.16	.071
Subjects Can Be Objects * Proportion Casemarked * Day=3	-0.10	0.17	.541
Subjects Can Be Objects * Proportion Casemarked * Day=4	-0.14	0.18	.432
Subjects Can Be Objects * Majority Word Order * Day=2	0.01	0.16	.970
Subjects Can Be Objects * Majority Word Order * Day=3	-0.14	0.17	.399
Subjects Can Be Objects * Majority Word Order * Day=4	-0.10	0.18	.567
Majority Word Order * Proportion Casemarked * Day=2	0.38	0.24	.111
Majority Word Order * Proportion Casemarked * Day=3	0.66	0.27	$.013^{*}$
Majority Word Order * Proportion Casemarked * Day=4	1.43	0.30	$< .001^{***}$
Subjects Can Be Objects * Majority Word Order * Proportion Casemarked	0.25	0.36	.487
Subjects Can Be Objects * Majority Word Order * Proportion Casemarked * Day=2	-0.30	0.31	.330
Subjects Can Be Objects * Majority Word Order * Proportion Casemarked * Day=3	-0.31	0.33	.349
Subjects Can Be Objects * Majority Word Order * Proportion Casemarked * Day=4	-1.30	0.36	$< .001^{***}$