You say yes, I say no: Investigating the link between meaning and form in response particles

Mora Maldonado, Jennifer Culbertson

Maldonado, Mora & Culbertson, Jennifer. 2023. *You say yes, I say no*. Investigating the link between meaning and form in response particles. *Glossa: a journal of general linguistics* 8(1). pp. 1–41. DOI: https://doi.org/10.16995/glossa.9185

You say yes, I say no. Investigating the link between meaning and form in response particles

Mora Maldonado, Laboratoire de Linguistique de Nantes, CNRS, Nantes Université, FR, mora.maldonado@univ-nantes.fr

Jennifer Culbertson, Centre for Language Evolution, University of Edinburgh, UK, jennifer.culbertson@ed.ac.uk

Response particles, like English 'yes' and 'no', are used to respond to polar questions or assertions and are found in all languages. However, the number of particles and the specific meanings they convey vary across languages. For example, in some languages particles mainly convey whether the response itself is positive or negative, while in others they convey whether the response is agreeing or disagreeing with previous discourse. Further, some languages have two response particles, while others have three, or even four. Recent work suggests that how meanings tend to be mapped to forms cross-linguistically might nevertheless be constrained. Roelofsen and Farkas (2015) suggest that indicating disagreement with a negative question or assertion (e.g., A: 'Ally doesn't eat meat.' B: 'Yes, he does.') is more marked than indicating agreement with a positive assertion (e.g., A: 'Ally eats meat.' B: 'Yes, he does'.). This difference in semantic markedness is argued to lead to a difference in form: more marked meanings are mapped to more specialized forms. Here we investigate this hypothesis in a series of behavioral experiments. Across our experiments, we find that participants are indeed sensitive to the differences in meaning that particles can convey. However, not all of the differences implicated by the hierarchy hypothesized in Roelofsen and Farkas (2015) are supported by our results, and we find evidence highlighting an unexpected special role for Positive Agreement—the least marked meaning.

1 Introduction

Response particles are words like English 'yes' and 'no', commonly used to respond to polar questions or assertions as shown in **Table 1**. Such particles are found in all languages, but the number of them and the specific meanings they convey vary across languages (see Pope 1976; Jones 1999; Holmberg 2013: among others). Nevertheless, previous work has argued that two types of features universally govern the use of these particles (Pope 1976; Roelofsen & Farkas 2015; see also Kramer & Rawlins 2009; Krifka 2013; Holmberg 2013; Pasquereau 2020b, for alternative, not feature-based, accounts). These features capture two dimensions of meaning that vary across the pairs in **Table 1**: first, whether the response itself has a positive or negative polarity (i.e., its *absolute* polarity), and second, whether the response agrees or disagrees with the preceding discourse (i.e., its polarity *relative* to an antecedent). Following previous work, we will refer to absolute features as Positive and Negative, and to relative features as Agreement and Disagreement.[1]

	Agreement	**Disagreement**
Positive	A: Ally eats meat.	A: Ally doesn't eat meat.
	B: Yes, he does.	B: Yes (No), he does.
Negative	A: Ally doesn't eat meat.	A: Ally eats meat.
	B: No (Yes), he doesn't.	B: No, he doesn't.

Table 1: English response particles as a function of preceding discourse, illustrating preferred usage for each of the two dimensions of meaning: absolute polarity in the response (Positive/Negative) and polarity relative to previous discourse (Agreement/Disagreement). Particles in parentheses are possible, though not preferred, for the relevant context.

Languages can in principle use a set of particles to convey only one of these feature contrasts, leaving the other contrast underspecified (or neutralized). For example, Japanese particles have traditionally been described as encoding relative polarity, with 'hai' and 'iie' analysed as conveying Agreement and Disagreement respectively (Pope 1976; Sadock & Zwicky 1985). English is often described as mainly conveying absolute polarity: 'yes' is used for Positive answers, and 'no' for Negative answers. However, English response particles appear to also be sensitive to the polarity of the preceding discourse or antecedent: given a negative assertion (e.g., 'Ally doesn't eat meat'), 'yes' can also sometimes convey agreement, as in Negative Agreement contexts, and 'no' can sometimes be used to convey disagreement, as in Positive Disagreement contexts (Goodhue & Wagner 2018; Repp et al. 2019).

[1] Relative features are also known as [AGREE/REVERSE] and absolute features as [+/–] (Roelofsen & Farkas 2015). Instead of Positive Disagreement, in this work one would talk about [REVERSE, +].

Other languages have a larger set of particles, in which case one particle is often used to convey not an individual feature, like Positive or Negative polarity, but a specific feature *combination*, i.e., a particle that is restricted to one cell in **Table 1**. For example, French has 'oui' and 'non' which have been analyzed in a similar way to English 'yes' and 'no'. However, a third particle 'si' also exists alongside these, and has been argued to express Positive Disagreement. Under this analysis, 'si' is a specialised particle used only when the response disagrees with a preceding negative question or assertion. Similar three-way particle systems, with a third particle conveying Positive Disagreement are found in a number of languages, like German, Dutch, Norwegian, Slovene, Farsi, and some varieties of Arabic (for a more complete list see Moser 2019). All together, these three types of systems (e.g., Japanese-, English- and French-type systems) represent the majority of documented response particle systems (Pope 1976).

Most theories of response particles have focused on the behaviour of these particles in the different contexts where they can appear (especially in scenarios where they are interchangeable e.g., English 'yes' and 'no' in Negative Agreement contexts)(Kramer & Rawlins 2009; Krifka 2013; Farkas & Roelofsen 2019; Pasquereau 2020b). Only some of these theories have also attempted to account for the constrained variability in response particle systems across languages (Pope 1976; Roelofsen & Farkas 2015; Farkas & Roelofsen 2019). These proposals share the idea that an important explanatory factor behind cross-linguistic patterns is the existence of asymmetries between the four possible meanings in this semantic space, i.e., Positive Agreement, Negative Agreement, Positive Disagreement, and Negative Disagreement.

In what follows, we will specifically focus on these last set of theories, as they establish an explicit connection between meaning and form. As a consequence, these theories make predictions about how response particle systems could partition the meaning space across languages. We leave aside other proposals, whose predictions regarding possible particle systems are less straightforward.[2]

Pope (1976) and Roelofsen & Farkas (2015) (see Farkas & Roelofsen 2019, for a posterior proposal) suggest that possible particle meanings differ in their semantic 'markedness', a notion which conveys that there are semantic or pragmatic differences of some sort between them. These differences might amount to, for example, relative processing cost, or richness in (or complexity

[2] For example, an alternative to Roelofsen & Farkas (2015) is given in Krifka (2013). In this account, response particles are anaphors that pick up a proposition introduced by their antecedent, and either negate it or affirm it. Importantly, negative antecedents (e.g., 'Ally doesn't eat meat') introduce both positive and negative discourse referents. So, whenever there is a negative antecedent, there is a choice with respect to which discourse referent should be affirmed or negated. By default, positive antecedents are assumed to be more salient than negative antecedents. As a result, in Negative Agreement scenarios, it would be preferable to use the particle 'no' to negate the positive discourse referent than to use the particle 'yes' to affirm the negative one. This accounts for preference patterns in languages like English. However, it doesn't make any predictions about what types of particle systems should be more or less likely in the typology (see Farkas & Roelofsen 2019, for discussion about this point).

of) semantic content. Regardless of the exact nature of the difference, in these accounts, the more marked the meaning is, the stronger the pressure is for it to be unambiguously expressed. As a result, marked meanings are most likely to be conveyed by specialised particles, like French 'si'. This is reminiscent of a more general idea in linguistics which links the presence of grammatical marking to semantic markedness, predictability, and/or frequency (e.g., see Haspelmath 2021). To cash out the idea in the case of response particles, Roelofsen & Farkas (2015) hypothesize a markedness hierarchy among possible meanings:

(1) Positive Agreement < Negative Disagreement < Negative Agreement < Positive Disagreement (where '<' indicates less marked).

The markedness of meanings is mainly grounded in the claim that there is a natural association between Negative polarity and Disagreement, on the one hand, and Positive polarity and Agreement, on the other. Negative polarity and Disagreement have in common the rejection or denial of a given proposition; Positive polarity and Agreement have in common its acceptance or affirmation. The two types of features of course differ in that for relative features this semantic content (denial or affirmation) may not be formally expressed in the response. This observation is argued to lead to two asymmetries, resulting in the markedness hierarchy in (1).

First, response particles which involve incongruity of absolute and relative features are assumed to be more marked. This results in an asymmetry between Positive Agreement and Negative Disagreement, which are congruent, and Negative Agreement and Positive Disagreement, which are incongruent. It's important to note that feature (in)congruence directly relates to whether the antecedent of the response has negative polarity or not: incongruent responses agree or disagree with a negative antecedent (Positive Disagreement and Negative Agreement). Second, meanings which involve computing negation or denial are assumed to be marked relative to those which involve affirmation. This markedness of negation is supported by various sources of evidence. For example, stand-alone negative statements are typically less informative (Horn 1989), less frequent, and generally harder to process than positive ones (see Tian & Breheny 2016, for a review). Perhaps as a result, negation is always overtly expressed morphosyntactically (Greenberg 1966). Similar patterns have been found for false statements relative to true ones (Trabasso et al. 1971). The markedness of negation leads straightforwardly to the hypothetized markedness contrast between Positive Agreement and Negative Disagreement: even though both are congruent, the latter involves negation (and denial).

Importantly, both Pope (1976) and Roelofsen & Farkas (2015) argue for an additional distinction between Negative Agreement and Positive Disagreement. In particular, these theories posit that Positive Disagreement is the most marked meaning in the hierarchy. This is less straightforward, since both involve negation/denial. However, under both theories, Positive Disagreement is argued to introduce a contrast/mismatch between the negative preceding discourse and the positive response. This is not the case for Negative Agreement, where the

response shares the negative polarity of the preceding discourse. Another way of thinking about this is that Positive Disagreement has something in common with double negation—known to be infrequent and difficult to process (Bellugi 1967; De Swart 2009; Blanchette 2015, among others). Positive Disagreement responses require the denial of a negative statement, which involves computing two consecutive negations.

In a recent study, Noveck et al. (2021) provide some behavioural evidence for the markedness of Positive Disagreement meanings relative to others. They show that both child and adult French speakers take longer to provide 'si' responses to negative interrogatives compared to responses involving 'oui' or 'non' answers. This finding suggests that, at least in production or response planning, Positive Disagreement is indeed marked. Importantly, the authors argue that this processing difficulty does *not* obviously follow from low frequency of usage of 'si' (in the Lexique2 database, New et al. 2004: 'oui': 3207 instances per million; 'si': 2107; 'non': 4040). Although 'si' is less frequent than the other two particles, the difference is perhaps not substantial enough to drive a meaningful markedness effect.

This finding notwithstanding, one of the main sources of support Pope (1976) and Roelofsen & Farkas (2015) cite for their theory is the very existence of three-way particles systems like French. As mentioned above, these systems have been analysed as having two general particles which refer to individual features like Negative and Positive polarity, plus a third specialized particle conveying Positive Disagreement, the most marked meaning according to the hierarchy in (1). This type of system is predicted under the assumption that there is a connection between the markedness of a meaning and its likelihood of being expressed using a specialized particle.

By contrast, their theory predicts that no language should have a specialized particle to express e.g., Positive Agreement, the least marked meaning. In other words, according to Roelofsen & Farkas (2015), languages might have only particles that express individual feature values, like Positive and Negative polarity (as in English), or Agreement and Disagreement (as in Japanese), but if a language has particles which convey specific meanings, i.e., feature combinations, these will convey meanings which are high on the markedness hierarchy. This is confirmed in the survey of languages reported in Roelofsen & Farkas (2015) and Farkas & Roelofsen (2019): there are languages with a specialized Positive Disagreement particle, and at least one which they analyze as having a specialized Negative Agreement particle (the Ethiopian language Soddo, as described in Leslau 1962), but apparently no languages with a specialized particle expressing Positive Agreement or Negative Disagreement.[3]

[3] There is also some evidence that the strategy of expressing more marked meanings (e.g., Positive Disagreement) using more marked forms may apply beyond the lexical level: in languages where there is no specific response particle to express feature combinations such as Positive Disagreement, positive responses to negative statement are often marked through prosodic or gestural patterns, or even repetition of particles (e.g., Goodhue & Wagner 2018 for English and González-Fuente et al. 2015 for Catalan and Russian).

However, the sample of languages in Roelofsen & Farkas (2015) is very small (just a handful), and in a much larger typological sample, Moser (2019) argues that a dedicated particle to express Positive Disagreement is largely a Western European phenomenon. On top of that, while there is evidence for this specific analysis in German (Claus et al. 2017), in a system like French, it is not entirely straightforward to see why 'si' should be analyzed as a dedicated Positive Disagreement particle, but 'oui' should be treated as a general particle expressing the Positive feature, rather than Positive Agreement specifically. Such an analysis is dependent on the possibility of using 'oui' in both Positive Agreement and Positive Disagreement contexts; in other words, if 'oui' were only found in Positive Agreement contexts, then it would have to be analysed as exactly the kind of particle Roelofsen & Farkas (2015) argue does not exist. Indeed, a number of studies on French suggest that 'oui' *is* mainly used in Positive Agreement contexts (Diller 1984; Takagaki 2014; Noveck et al. 2021, but see Pasquereau 2020a for an account of how the distribution of French particles depends on the inclusion of the prejacent in the response). Finally, the justification provided for the hierarchy in (1) is not entirely intuitive. For example, one could easily imagine that Negative Disagreement might be the most marked meaning since it combines what Roelofsen & Farkas (2015) argue are the two most marked features (see also Goodhue & Wagner 2018). Instead, according to the hierarchy it is less marked than Negative Agreement which only conveys one marked feature. In sum, there are a number of reasons to test this theory further.

More generally, linguistic theories based on typological data alone (cf. Bellugi 1967) are rarely enough to tell us why certain patterns—e.g., particular kinds of particle systems—appear more commonly than others. Typological distributions are driven by many factors, and do not by themselves support causal links with hypothesized properties of our cognitive or linguistic system, like the markedness hierarchy in (1) (Culbertson 2023). Here we present a series of experiments aimed at providing additional evidence for the psychological reality of this hierarchy, and the potentially special status of particles expressing Positive Disagreement. One challenge for doing so lies in operationalizing the notion of markedness. As mentioned above, markedness in meaning is a vague concept, often associated with frequency or predictability. In much literature connecting markedness of meanings to markedness of forms, the latter is associated with length: more marked meanings are expressed using longer words, or are not lexicalized and thus have to be expressed using phrases (as in some attested particle systems, see Footnote 3). However, in Roelofsen & Farkas (2015), markedness is operationalized in terms of a *specialized form*, with restricted domain of application.

In Experiments 1 and 2 we cash out the notion of markedness in these two different ways—length and restricted distribution—and test whether English-speaking participants learning a novel three-way particle system prefer to align more marked particles with more marked meanings (according to the hierarchy in 1). We then present a final experiment (Experiment 3)

aimed at assessing whether English speakers are sensitive to some of the markedness contrasts behind the hierarchy in (1) (i.e., feature (in)congruence), when they interpret meanings.

2 Experiment 1

In this first experiment we derive a corollary prediction from the theories discussed above, namely that learners acquiring a new particle system should naturally align more marked meanings with more complex forms. This prediction extends proposals such as Pope (1976) and Roelofsen & Farkas (2015) in two ways. First, it makes explicit the linking mechanism by which the hierarchy in (1) finds its way into the typology: through a cognitive bias in learning. This is a common linking mechanism (either implicitly assumed, or explicitly posited) in linguistic theory, and it has been tested using the experimental methods we use here (i.e., artificial language learning, see Culbertson 2023, for a review). Briefly, the idea here is that typological patterns reflect biased learning: systems that are hard (or impossible) to learn are less likely to be passed on from generation to generation with high fidelity, and are therefore rare (or unattested). Second, our prediction equates the notion of a specialized particle with formal complexity; in particular, with word length. As mentioned above, this is related to a broader class of linguistic phenomena in which meanings that are surprising, unexpected, or unusual in some way are expressed using more linguistic content (Haspelmath 2021).

Participants in this experiment were trained on a language with two response particles conveying all four possible meanings: Positive Agreement, Negative Disagreement, Negative Agreement, and Positive Disagreement. Critically, one of the two particles in each condition randomly varied between a full and clipped form. It is now well established that learners tend to regularize inconsistent variation of this kind during laboratory learning (e.g., Hudson Kam & Newport 2009; Reali & Griffiths 2009; Smith & Wonnacott 2010; Ferdinand et al. 2019; Saldana et al. 2021). In some cases this is driven by a preference for one variant over the other (e.g., Fedzechkina et al. 2012; Culbertson et al. 2012; Culbertson & Newport 2015; Culbertson et al. 2020), one mapping between variants and meanings over other (e.g., Smith & Wonnacott 2010), or based on word frequency/predictability or other notion of markedness (Kanwal et al. 2017a; b; Kurumada & Grimm 2019). Here, we test whether learners will condition the variation between the full and clipped form of the particle on which meaning is being expressed. The prediction we test is that learners will implicitly expect more marked meanings to be conveyed by more marked forms, and therefore will preferentially use the longer full form to convey meanings that are relatively more marked according to the hierarchy in (1). Each of the four conditions allows us to compare a pair of meanings which differ in markedness, as described in detail below. The experiment reported here, including all exclusion criteria and statistical analyses, was preregistered: https://osf.io/gtpf6.

2.1 Methods

2.1.1 Stimuli

Participants were trained on two novel response particles, used as responses to statements like those in **Table 2**. The organization of the novel particles into systems varied across participants as shown in **Table 3**.

	Agreement	**Disagreement**
Positive	A: The square is red.	A: The square is not red.
	B: *Yes*, it is.	B: *Yes (no)*, it is.
Negative	A: The square is not red.	A: The square is red.
	B: *No (yes)*, it's not	B: *No*, it's not.

Table 2: Contexts for the use of response particles in Experiment 1. For comparison, we give English response particles (with possible but less preferred options in parentheses). In the experiment itself, these were replaced by the forms from the relevant system in Table 3.

Condition 1a			**Condition 1b**		
	Agreement	Disagreement		Agreement	Disagreement
Positive	*va(dof)*	*va(dof)*	Positive	*va(dof)*	*nish*
Negative	*nish*	*nish*	Negative	*va(dof)*	*nish*
Condition 1c			**Condition 1d**		
	Agreement	Disagreement		Agreement	Disagreement
Positive	*vash*	*mi(dof)*	Positive	*vash*	*vash*
Negative	*vash*	*mi(dof)*	Negative	*mi(dof)*	*mi(dof)*

Table 3: Artificial response particle systems in the four conditions of Experiment 1. In the top two conditions, the particles convey Positive and Negative features. In the bottom two, they convey Agreement and Disagreement. One particle varies in length and is used equally frequently to express the two critical meanings.

In two conditions (Conditions 1a and 1d), these two particles conveyed the two absolute features, Positive and Negative (similar to English). For example, in Condition 1a, one particle was used in both Positive Agreement and Disagreement contexts and the second particle was used in both Negative Agreement and Disagreement contexts. In the other two conditions (Conditions 1b and 1c), the two particles conveyed the two relative features Agreement and Disagreement, (similar to Japanese, though note that English particles *can* also convey relative features as

shown in **Table 2**). For example, in Condition 1b, one particle was used in both Positive and Negative Agreement contexts and the second particle was used in both Positive and Negative Disagreement contexts.

In all four conditions, one of the response words did not vary (*nish* in Conditions 1a and 1b, *vash* in Conditions 1c and 1d, designed to be somewhat similar to English 'no' and 'yes' respectively to facilitate learning). Importantly, the other particle word varied between two alternating forms, a longer full form and a shorter clipped form (*va(dof)* in Conditions 1a and 1b, *mi(dof)* in Conditions 1c and 1d). Use of the long and short forms for these two critical meanings was random (both forms were used exactly half of the time for each meaning). Other than the novel particles, the verbal stimuli were in English.

As mentioned above, response particles were presented in the context of picture descriptions as in **Table 2**. Pictures consisted of two shapes (from the set {square, circle, diamond, triangle, heart, star}), each of a different color (from the set {red, blue, orange, yellow, pink, purple, brown}). Speaker A described the color of one of the shapes in the picture. The description was either correct or incorrect. Speaker B then provided a response, using one of the response particles and the continuation 'it is' or 'it's not'.

2.1.2 Procedure

In the training phase, participants were first informed that they would see pictures of simple shapes, and then a description of the picture given by a speaker (Speaker A), which might be correct or incorrect. They were told that they would then see a response to this description by a second speaker (Speaker B). This response would contain unfamiliar words. Participants were instructed to learn the meanings of the words so that they could use them later.

During training, each context was presented 6 times, for a total of 24 trials, in random order per participant. The description given by Speaker A always referred to the color of one of the two shapes in the picture. Depending on the context, it involved a positive or a negative description. The description was correct half the time (in Agreement contexts) and incorrect half the time (in Disagreement contexts). As described above in the critical contexts for each condition, half of the responses contained the short clipped form of the relevant particle and half the long form. Each trial proceeded as follows: a picture appeared (2s), followed by Speaker A's description (3s), and then a Speaker B's response (3s). Then the response disappeared and participants had to choose which of the three possible particle forms in the language was just used by clicking on the appropriate button. If they were correct, the button turned green and the trial progressed after 1s; if they were incorrect, the button turned red and the trial progressed after 3s. An example trial is shown in **Figure 1**.

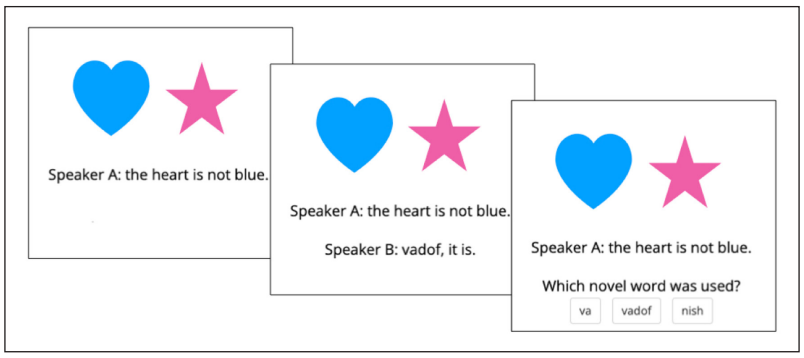

Figure 1: Training trial procedure in Experiment 1 (Condition 1a; Positive Disagreement context).

In the testing phase, participants were first told that they would see pictures and descriptions as before, but this time they had to fill in the response given by Speaker B. During testing, each context was presented 4 times, for a total of 16 trials, in random order per participant. On each trial, a picture appeared (1s), followed by a description (1s), and then three text input boxes appeared. Participants had to type a novel particle (in the first box) along with the sentence continuation 'it is' or 'it's not' (in the second and third box). Participants were required to complete all input boxes. If they failed to fill in all three boxed they were warned 'Please fill all the fields'. If they typed a continuation that failed to use the correct polarity (e.g.,' not' in the third box on a Positive Agreement or Positive Disagreement context, or 'is' in the third box for a Negative Agreement or Negative Disagreement context), they were warned 'Please check your answer. Is the shape the color Speaker A says it is?'. The trial advanced only once these criteria were satisfied. Critical testing trials were those involving the two meanings conveyed by the particle which varied in length (e.g., Positive Agreement and Positive Disagreement contexts for Condition 1a). An example trial is shown in **Figure 2**.

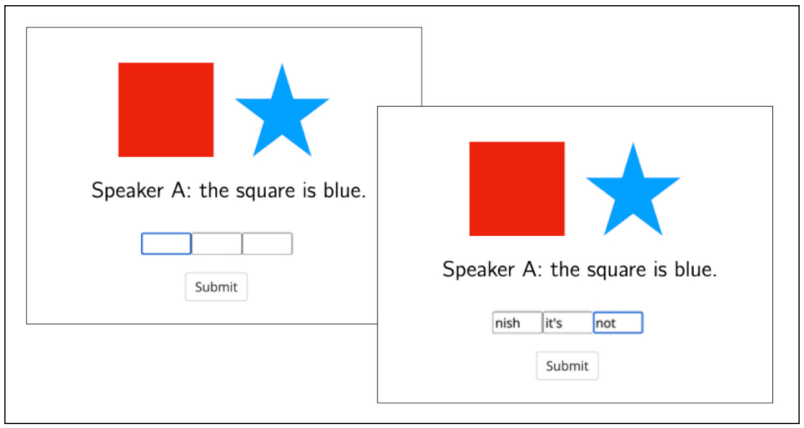

Figure 2: Testing trial procedure in Experiment 1 (Condition 1a; Negative Disagreement context).

2.1.3 Participants

Participants were recruited through Amazon Mechanical Turk or Prolific and gave informed consent.[4] They received 2.50 USD as compensation on Mechanical Turk and 3.00 GBP on Prolific (due to a rise in the living wage during the study).

Participants were self-reported native English speakers. In the case of Prolific we used the available pre-screening criterion to recruit only monolingual English speakers. In the case of Mechanical Turk, we relied on participants self-reports. English speakers are a good population to test in this case, since their native language has what have been analysed as underspecified or general particles expressing (for the most part) positive or negative polarity. Critically, English does not have a three-way particle system with specialized particle expressing a specific feature combination—the type of particles we are interested in here. Further, two of the four conditions tested here are similar to English in the sense of having particles that generally refer to absolute features, and two are distinct from English in having particles that generally refer to relative features.

We recruited a total of 328 participants (Condition 1a: 77; Condition 1b: 67; Condition 1c: 125; Condition 1d: 61). We excluded Mechanical Turk participants who were self-reported speakers of a language with a three-way particle system (Condition 1a: 8; Condition 1b: 17). Per our pre-registration, we further excluded participants whose accuracy on testing trials was less than 80% (Condition 1a: 24; Condition 1b: 20; Condition 1c: 77; Condition 1d: 13), where accuracy was defined as typing a response that has an edit distance of 1 or less to the correct particle for a given trial (both the full and clipped forms were treated as correct for the two critical meanings). The data of the remaining 188 participants were analyzed (Condition 1a: 45; Condition 1b: 47; Condition 1c: 48; Condition 1d: 48).

2.2 Results

Results from Experiment 1 are shown in **Figure 3**. Data were analyzed using mixed effects logistic regression (here and throughout all analyses use the lme4 package in R; Bates 2010; R Core Team 2020). The data of interest are responses from trials involving the critical particle which alternated in length and could be used in two contexts for each condition. We analyse each condition separately. For each condition, we fit two models, one null model with only an intercept term and a second model with meaning as a fixed effect (two levels, always the less marked meaning and the more marked meaning for each condition, sum coded). Both models for each condition included by-participant random intercepts, and random by-participant slopes for meaning. We compared each pair of models using a Likelihood Ratio test. For Condition 1a,

[4] We switched from Mechanical Turk to Prolific during the time in which this study was conducted following a decline in quality of participants we noticed on Mechanical Turk during the pandemic. Prolific also allows pre-screening of participants according to self-reported language knowledge, which we use here.

the test revealed that the more complex model, including meaning, provided a better fit to the data ($\beta = 0.80$, $SE = 0.30$, $\chi^2(1) = 7.10$, $p = 0.008$). For Condition 1b, including meaning also provided a better fit to the data ($\beta = 1.04$, $SE = 0.26$, $\chi^2(1) = 15.76$, $p < 0.001$). For Condition 1c, including meaning failed to provide a better fit to the data ($\beta = –0.25$, $SE = 0.31$, $\chi^2(1) = 0.60$, $p = 0.44$). Finally, for Condition 1d, including meaning again failed to provide a better fit to the data ($\beta = 0.40$, $SE = 0.31$, $\chi^2(1) = 1.73$, $p = 0.19$).[5]

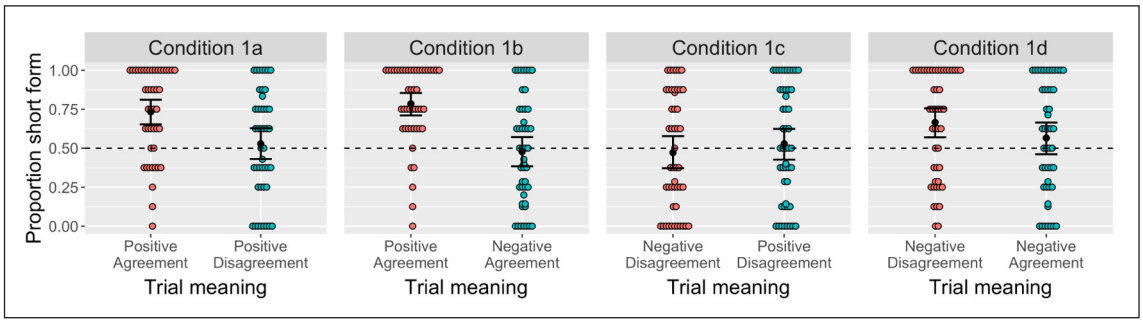

Figure 3: Proportion of short form usage for critical meanings in the four conditions of Experiment 1. Dots are individual participants, error bars are 95% bootstrapped confidence intervals.

2.3 Discussion

In Experiment 1, we trained participants on a new two-way response particle system in which one of the particles alternated randomly between a long and short form (e.g., *va* vs. *vadof*). The systems varied in terms of which two meanings were conveyed by the alternating particle. Our aim was to determine whether learners acquiring such a system would expect the longer form of the alternating particle to be mapped to a more marked meaning according to the markedness hierarchy in (1). This kind of relationship between length and meaning markedness is well-attested in language, and may hold in some form in attested response particle systems. Our results provide some evidence that learners do indeed condition particle length on meaning. In Conditions 1a and 1b, participants used the shorter form more for Positive Agreement, and the longer form more for the more marked meaning—Positive Disagreement in Condition 1a and Negative Agreement in Condition 1b. This is consistent with the learning predictions we derived from the markedness hierarchy in (1). However, in Conditions 1c and 1d, we failed to

[5] There is a numeric trend in the predicted direction for Condition 1d (effect size of 0.40). To confirm the failure to find a significant effect of meaning in this condition, we ran a replication, with 60 additional participants. We again failed to find any evidence that participants conditioned the use of the long and short forms on meaning ($\beta = –0.09$, $SE = 0.28$, $\chi^2(1) = 0.10$, $p = 0.75$). In this case, there was no trend in either direction, and numerically more long form responses for the *less marked* Negative Disagreement meaning.

find evidence that participants' use of the long or short form was conditioned on the meaning to be conveyed.

One potential interpretation of these results is that learners' behavior in Experiment 1 points toward a special role for *Positive Agreement*. Where Positive Agreement was among the critical meanings, participants preferred to map the *short form* to this meaning. Where it was absent, participants showed no preference. The pattern of results in Conditions 1a and 1b may thus reflect participants' preference to use short forms with Positive Agreement (rather than using long forms with e.g., Positive Disagreement, as we expected). This could be due to the fact that this meaning is the least marked in the hierarchy. However, given that we did not find any other evidence in support of the hierarchy (i.e., use of the shorter form with Negative Disagreement, also relatively unmarked) it may simply be that participants find Positive Agreement easy to process or learn. Indeed, participants generally had the highest accuracy in Positive Agreement contexts (see Footnote 6). This means that participants were not just more likely to use the short form for Positive Agreement contexts, they were also less likely to use an incorrect particle in these contexts compared to any other across conditions. Therefore, it is possible that our results point not to any sensitivity to the complex of markedness relations in (1) during learning, but instead to the special status of Positive Agreement, e.g. as least marked, semantically or otherwise.

However, it is also worth considering a few other possibilities. First, Condition 1c was very difficult for participants to learn (there were many more exclusions for particle accuracy in this condition). Participants' failure to condition short vs. long form use on meaning in this condition may, at least in part, be due to the difficulty of mapping the particles to meanings in this condition.[6] Particle accuracy in Condition 1d was similar to Conditions 1a and 1b, but the lack of sensitivity to markedness in this condition might be due to the fact that these two

[6] We conducted a mixed-effects logistic regression analysis of all pre-exclusion data, including fixed effects of condition (baseline level: Condition 1a, treatment coded), meaning (baseline level: Positive Agreement, treatment coded), and their interaction. Note that this analysis was not pre-registered, as we did not have any hypothesis about which of these systems would be easier to learn overall. The analysis revealed a significant difference between Condition 1a and Condition 1c ($\beta = -5.29$, $SE = 1.22$, $p < 0.001$). No significant differences were found between Condition 1a and any other condition overall. Additionally, accuracy for Positive Agreement was higher than for all other meanings in Condition 1a (all $\beta = -4$, all $p < 0.001$). Significant interaction terms showed that the differences between meanings in some cases differed in other conditions, but there was no evidence that any particular meaning was learned worse in Condition 1c. In a separate experiment not reported here, we also ran a version Condition 1c with a different set of particles. As noted, forms used in Experiment 1 for Negative polarity and/or Disagreement were mapped to nonsense particles somewhat resembling English 'no', and Positive polarity and/or Agreement meanings were mapped to nonsense particles somewhat resembling English 'yes'. In this additional version of Condition 1c, the Disagreement meanings were mapped to *va(dof)*, and Agreement meanings to *jesh*. This change does not impact the results, ($N = 29$, again after substantial exclusions, 37 total, based on accuracy) which showed no significant effect of meaning on long/short particle use ($\beta = 0.25$, $SE = 0.37$, $\chi^2(1) = 0.48$, $p = 0.49$).

meanings are more similar to one another in terms of markedness (i.e., they are adjacent on the hierarchy).

Second, it might be that not every semantic property that contributes to the markedness hierarchy is equally important *during learning*. Learners may be sensitive to differences in markedness that result from whether or not one of the critical meanings involved Negative polarity or Disagreement, but not to differences in markedness that are due to feature (in)congruence. In Conditions 1a and 1b, the critical meanings were Positive Agreement and either Positive Disagreement (1a) or Negative Agreement (1b). The latter two involve both Disagreement/ Negative polarity and feature incongruence, the former involves neither, and here learners condition the long and short forms as predicted. By contrast, in Conditions 1c and 1d, one of the critical meanings involves congruent features (Negative Disagreement) and the other incongruent features (Positive Disagreement in 1c or Negative Agreement in 1d)—and here learners do not condition the long and short forms as predicted. Under this interpretation, learners would thus be sensitive to markedness based on specific features, but not based on feature (in)congruence.

Along similar lines, it could also be that the special role of Positive Agreement we observe is driven by its contrast on *both these dimensions* with the alternative meanings in Conditions 1a and 1b. In other words, in these two conditions, the critical meanings differ both in negativity *and* feature (in)congruence, whereas in Conditions 1c and 1d, the critical meanings differ only in feature (in)congruence. Perhaps in the context of this experiment, learners are sensitive to meaning contrasts only when they involve differences in more than one dimension.[7] In any case, this would leave unexplained the mechanism through which the rich set of markedness contrasts encoded by the hierarchy impact the typology of these particles systems. In particular, here we found no evidence that differences driven by feature (in)congruence alone impact learning. Further, there are a number of other cases in the literature in which learners *have* shown sensitivity to shared features—similar to feature congruence—in artificial language learning experiments (e.g., Maldonado & Culbertson 2020; Saldana et al. 2022). We return to this issue in Experiment 3 by looking for evidence of sensitivity to feature (in)congruence in a different type of task.

Perhaps relatedly, it could be that Disagreement is perceived as very marked, regardless of whether it is Positive or Negative. This could have led to a failure to clearly distinguish the two critical meanings along the lines predicted in Condition 1c. Similarly, perhaps Negative polarity is perceived as very marked, regardless of whether it is Agreement or Disagreement. This could have led to a failure to distinguish the two critical meanings in Condition 1d. While this would be consistent with the general idea that participants are sensitive to the presence of Disagreement/Negative polarity but not feature (in)congruence, it would *not* be in line

[7] We thank an anonymous reviewer for this suggestion.

with the hierarchy, according to which Negative Disagreement is one of the least marked meanings.[8]

Given the mixed evidence we found in Experiment 1, it is worth exploring the possibility that complexity, or specifically length, is not the best way to operationalize the notion of specialized form put forward in Roelofsen & Farkas (2015) and Farkas & Roelofsen (2019). In Experiment 2, we use a different notion of specialized particle, and a different design to test the *ease of learning* of different particle systems (rather than regularization of variable input). We investigate three-way particle systems where one particle is specialized by virtue of having a more restricted distribution relative to the other particles in the system. This notion of specialized particle is arguably closer to the one described for languages such as French, German and other systems analyzed as having a specialized particle for Positive Disagreement responses (Roelofsen & Farkas 2015; Farkas & Roelofsen 2019; Moser 2019).[9]

3 Experiment 2

In Experiment 2, we test the prediction that learners will find it easier to learn a particle system in which a *specialized* particle is mapped to a more marked meaning. Here, by specialized particle we mean a particle whose use is restricted relative to the other particles in the system. Participants were trained on a new three-way particle system. The system had two general particles used to convey a single feature each (and thus compatible with more than one context), and a third specialized particle which was restricted in its use to one specific context or feature combination. A system with similar characteristics to this (in particular to the bottom-left system in **Table 4**) is found, for example, in German. Claus et al. (2017) find that the general particles 'ja' and 'nein' in German are, for a majority of speakers, *preferably* used to convey Agreement and Disagreement respectively, whereas the specialized particle 'doch' can *only* be used in Positive Disagreement contexts. As a result, 'doch' is preferred in Positive Disagreement contexts, but 'nein' is also possible.[10] Here, we manipulate between subjects the context in which

[8] Of course, the difference in markedness targeted by Condition 1c is the same as Condition 1b, and yet the pattern of results is clearly different. However, these two conditions may be problematic to compare given the evident difficulty participants had in learning the system in Condition 1c.

[9] In another experiment, not reported here, we tried a third operationalization of the notion of specialized particle by using an odd-one-out design: Participants were trained on a three-way particle system, where two of the particles orthographically resembled English 'yes' and 'no', and a third particle did not resemble either English response particle. We manipulate between subjects whether this special novel particle conveys Positive Disagreement or Positive Agreement meanings. The details of this experiment, including results, can be found in the OSF repository. To summarize, we found no difference between the two conditions, and therefore no support for sensitivity to the hierarchy.

[10] This is an oversimplified description of Claus et al.'s findings. Claus et al. (2017) find that there is a second, smaller group of German speakers for whom the particle 'nein' seem to express absolute rather than relative features (i.e. Negative rather than Disagreement). For these speakers, the particle 'nein' is preferred in Negative Agreement scenarios. However, this is in principle irrelevant for our purposes.

the specialized particle is used. We test the prediction that the systems in which this particle is used to convey Positive Disagreement—the most marked meaning according to the hierarchy in (1)—will be easier to learn than systems where the particle conveys (a) Positive Agreement or (b) Negative Agreement.[11] This experiment, including all exclusion criteria and statistical analyses was preregistered: https://osf.io/jxuwc.[12]

Positive Disagreement (+/–)			**Positive Agreement (+/–)**		
	Agreement	Disagreement		Agreement	Disagreement
Positive	*vaf*	*vaf /fub*	Positive	*vaf /fub*	*vaf*
Negative	*zog*	*zog*	Negative	*zog*	*zog*
Positive Disagreement (Agr./Dis.)			**Negative Agreement (Agr./Dis.)**		
	Agreement	Disagreement		Agreement	Disagreement
Positive	*vaf*	*zog/fub*	Positive	*vaf*	*zog*
Negative	*vaf*	*zog*	Negative	*vaf /fub*	*zog*

Table 4: Experiment 2 conditions. Top two conditions involve general particles expressing Positive (+) and Negative (–) features. Bottom two conditions involve general particles expressing Agreement (Agr.) and Disagreement (Dis.) features. In each condition, the critical meaning can be expressed either using one of the general particles or a specialized particle.

3.1 Methods

3.1.1 Stimuli

Participants were randomly assigned to learn one of four three-way particle systems. In two of the conditions, there were two underspecified general particles used to convey Positive and Negative polarity responses respectively. In the other two conditions, the general particles conveyed Agreement and Disagreement respectively. In all conditions, a third particle conveyed a feature combination, and therefore its use was restricted to a single context. In the first two conditions, we compare a specialized Positive Disagreement particle (as in **Table 4** top left) to a specialized Positive Agreement particle (as in **Table 4** top right). In the second two conditions, we compare a specialized Positive Disagreement particle (as in **Table 4** bottom left) to a specialized Negative Agreement particle (as in **Table 4** bottom right). As noted above, the Positive Disagreement (Agr./Dis.) condition has some similarity with German (Claus et al. 2017). The Positive Disagreement (+/–) condition is analogous to the analysis given by Farkas & Roelofsen (2019) for French,

[11] See Footnote 22 for discussion of why we did not test a system with a specialized particle conveying Negative Disagreement (the last logical possibility).

[12] The second set of conditions described below is not included in the original preregistration, but the predictions and analysis remain the same.

where the particles 'oui' and 'non' convey Positive and Negative features respectively, whereas 'si' is specialized for Positive Disagreement contexts. Note, however, that as far as we know, the French particle 'oui' is restricted to Positive Agreement contexts (i.e., cannot also be used in Positive Disagreement contexts). As a matter of fact, Pasquereau (2020a) suggests that the French particle 'non' is the only alternative form that can occur in Positive Disagreement contexts (though only under special circumstances).[13] All three particles were randomly chosen for each participant from the set {*fub, zog, vaf, muz*}. Note that here all particles are the same length, and all have English neighborhood density 1 (Storkel 2013).

Response particles were presented in the context of picture descriptions as in Experiment 1 (see **Table 2**). Pictures consisted of a single shape (from the set {square, circle, diamond, triangle, heart, star}), with a color (from the set {red, blue, orange, yellow, pink, purple, brown}). Speaker A described the color of the shape in the picture. The description was either correct or incorrect. Speaker B then provided a response, using one of the response particles. Note that here, to simplify the task, we did not use a continuation 'it is' or 'it's not' for the response (Goodhue & Wagner 2018; Pasquereau 2020a: see for examples of how the use of continuation or prejacent can affect the distribution of particle forms).

3.1.2 Procedure

The procedure was generally similar to Experiment 1, with two main differences. First, there were two phases: an initial phase in which participants were trained and tested on the two underspecified, general particles, and a second phase where participants were introduced to the third particle, and tested on the whole system. Second, the testing phases involved a forced-choice task (i.e. choosing between two particles) rather than typing response task.

3.1.2.1 Initial training and testing (general particles)

Participants were first told that they would see descriptions like 'The square is blue' provided by a speaker (Speaker A), then a picture would appear which might or might not match the description.[14] They were told that then another speaker (Speaker B) would provide a response using one of two unfamiliar words. Participants were told that these words 'are similar to *yes* and *no* in English.' They were instructed to try to learn when each of these words is used, so they could use them later on. During this pre-training phase, each context was used in 8 trials,

[13] We chose to use an underspecified, general particle instead since otherwise both Positive Agreement and Positive Disagreement could be analysed as specialized particles (cf. Roelofsen & Farkas 2015). We return to this issue in the General Discussion.

[14] We changed the order of presentation so that the description preceded the picture in order to better encourage the participants to form an expectation of the picture given the description. It was our intuition that this might better capture what is marked about Positive Disagreement.

for a total of 32 trials, in random order per participant. On each trial, a description appeared (1.5s), then a picture (2s), and finally a reply (2s). Then the reply disappeared and participants had to choose which of the two particles was just used by clicking on the appropriate button. If they chose correctly, the button turned green and the trial progressed after 1s. If they chose incorrectly, the button turned red, and the trial progressed after 3s. An example trial is shown in **Figure 4**.

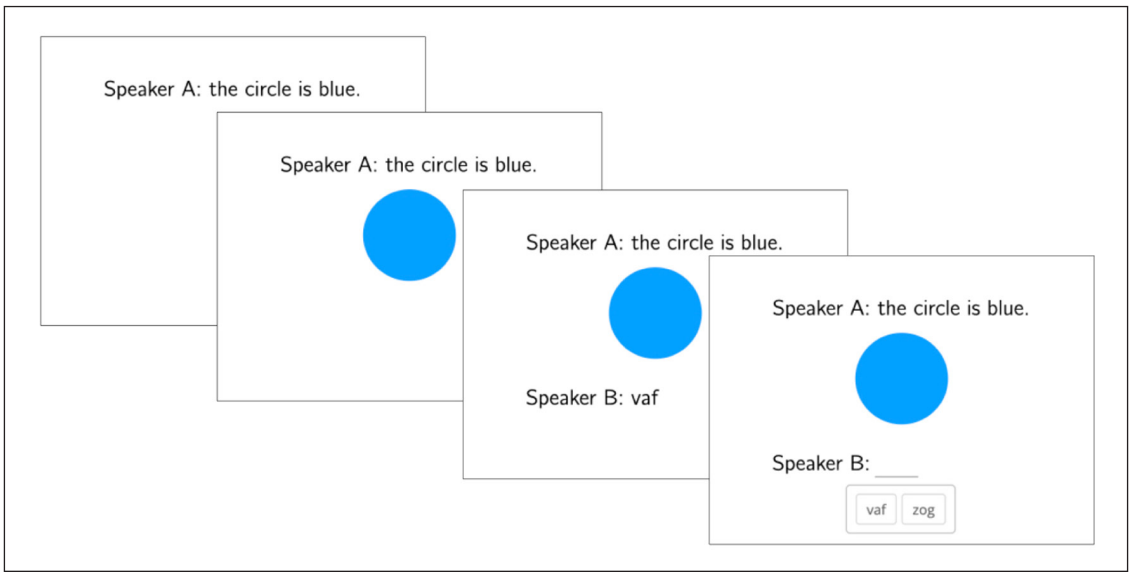

Figure 4: Training trial in Experiment 2 (Positive Agreement context).

After this initial training, participants were tested on their knowledge of the two general particles. During this phase, each context was used in 3 trials, for a total of 12 trials, presented in random order per participant. On each trial a description and picture appeared, and participants had to choose which of the two particles should be used by clicking on the appropriate button. If they chose correctly, the button turned green and the trial progressed after 1s. If they chose incorrectly, the button turned red, and the trial progressed after 3s. Note that, up to this point, participants in both conditions had received exactly the same training.

3.1.2.2 Critical training and testing (specific particles)

Participants were then told that they would learn about one more new word. They were told ‘this is a special word that can be used in a particular context, instead of one of the other words you’ve already learned. In other words, this language has not just words like ‘yes’ and ‘no’, but another word as well.’ They were instructed to try to learn when the special word could be used, so that they could use all three words they had learned in the right contexts.

During training, each context was used in 4 trials, plus an additional 4 trials for the critical meaning associated with the specialized particle in each condition. This made for a total of 20 trials, presented in random order per participant. On two of the 8 critical meaning trials, the general Positive particle was used (just as it had been during the general particle training phase), on the remaining 6 trials, the specialized particle was used. Training trial procedure was identical to the general particle training described above.

During testing, each context was used in 8 trials, plus an additional 4 trials for the critical meaning associated with the specialized particle in each condition. This made for a total of 48 trials, presented in random order per participant. Importantly, the 12 critical meaning trials were split between trials testing accuracy, and trials testing preference for the specialized particle over the alternative general particle. In the 6 critical accuracy trials, the foil response was always one of the two general particles, but the correct particle response could be either the specialized particle or the other general particle. In the 6 preference trials, the choice was between the specialized and the general particle that could occur in the same context—both correct, therefore testing whether participants preferred to use the specialized particle over a general particle for this meaning. **Table 5** summarizes the trial types for each of the four conditions.

	Accuracy (general)	**Accuracy (specialized)**	**Preference**
Positive Disagreement (+/–)	[+] vs. [–]	specialized vs. [–]	specialized vs. [+]
Positive Agreement (+/–)	[+] vs. [–]	specialized vs. [–]	specialized vs. [+]
Positive Disagreement (Agr./Dis.)	Agr. vs. Dis.	specialized vs. Dis.	specialized vs. Agr.
Negative Agreement (Agr./Dis.)	Agr. vs. Dis.	specialized vs. Agr.	specialized vs. Dis.

Table 5: Testing trial types for critical meanings across conditions in Experiment 2.

The remaining 36 non-critical meaning trials all tested accuracy; the correct answer was always one of the general particles and the foil was one of the other two particles in the language. These trials are summarized for the Positive Disagreement condition in **Table 5**. Testing trial procedure was identical to the general particle testing described above, expect there was no feedback provided. An example testing trial is shown in **Figure 5**.

3.1.3 Participants

All participants were recruited through the Prolific Academic online recruiting platform and gave informed consent. They received 3.00 GBP as compensation. We recruited a total of

228 participants, with the Prolific pre-screening criterion that they be monolingual English speakers. We excluded participants with an accuracy score of less than 66% (60) on the initial testing phase.[15] The data of the remaining 164 participants were analyzed (44 in the Positive Disagreement (+/–) condition, and 40 in the each in the remaining 3 conditions).

Figure 5: Testing trial in Experiment 2 (Positive Disagreement context).

3.2 Results

Recall that participants were trained on a three-particle system which involved two underspecified, general particles, and one specialized particle. Depending on the condition, the general particles were associated with Positive and Negative meanings, or with Agreement and Disagreement meanings. The specialized particle was used in either Positive Disagreement, Positive Agreement, or Negative Agreement contexts. Participants were tested on (a) how well they had learned the system, including the specialized particle (Accuracy trials), and on (b) whether they preferred to use the specialized particle over the underspecified alternative in critical contexts (Preference trials). Learners were expected to prefer systems where the specialized particle conveyed Positive Disagreement rather than Positive Agreement. An asymmetry between conditions was thus predicted both in learning and preference.

[15] Crucially, conditions do not differ at this point, the specialized particles have not yet been introduced, and therefore it is not problematic to exclude participants based on accuracy at this stage.

Figure 6 shows participants' performance on critical meanings in the final testing phase of the experiment, as a function of condition. Each facet of the plot shows one of the three types of trial described in **Table 5** above: (i) accuracy trials in which the correct response is a general particle, (ii) accuracy trials in which the correct response is the specialized particle, and (iii) preference trials, in which either the specialized particle or a general particle can be used. We first analyze accuracy trials, and then turn to preference trials. All data were analyzed using mixed effects logistic regression. Fixed effects were always sum-coded. Whenever possible we included a maximal random structure, following recommendations from Barr et al. (2013). P-values were obtained by a Likelihood Ratio test comparing the relevant model with a simpler one in which the relevant predictor was removed. Our confirmatory analyses involve comparisons between Positive Disagreement (+/–) and Positive Agreement (+/–) on the one hand, and Positive Disagreement (Agr./Dis.) and Negative Agreement (Agr./Dis.) on the other hand. In other words, we compare conditions which share the meaning of the general particles.

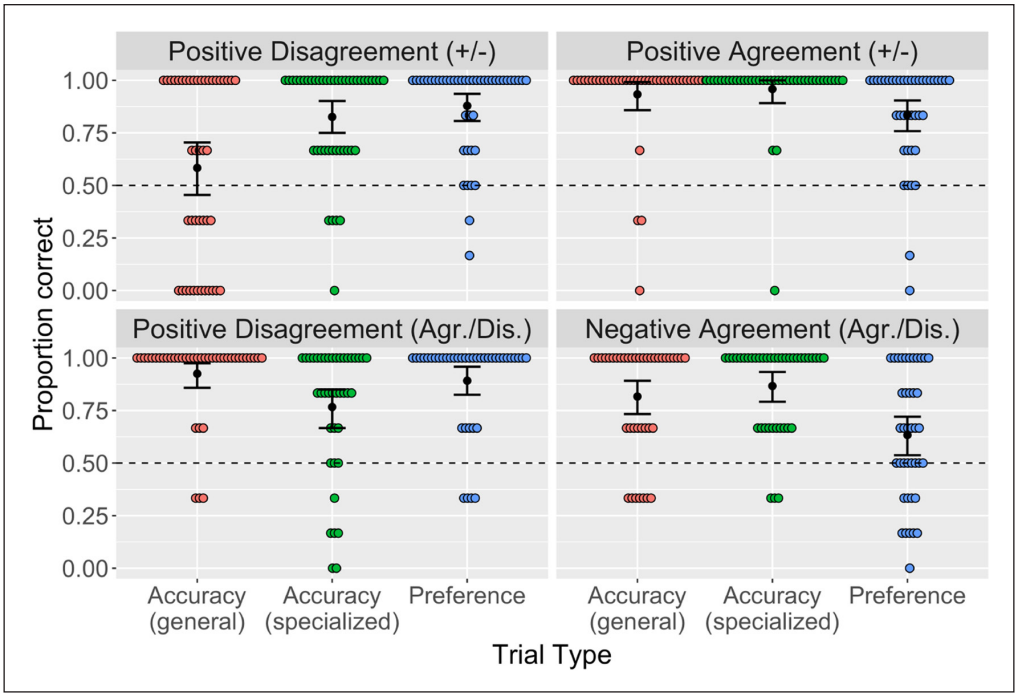

Figure 6: Responses for critical meanings across all conditions for accuracy and preference trials. Dots are individual participants, error bars are 95% bootstrapped confidence intervals. Note that for each condition there are two types of accuracy trials for critical meanings: one in which the correct response is the general particle, and the other in which the correct response is the specialized particle. For preference trials, either response is correct, but here we are treating the specialized particle responses as 1 and the general particle responses as 0. Therefore the axis for these trials represents proportion choice of the specialized particle. See Table 5 for detailed description of trials types for each condition.

We first test whether the overall accuracy of the entire system differed across the two relevant conditions (i.e., including all four meanings, not just the critical ones). We found no evidence for an effect of condition in either comparison (+/–: $\beta = 0.002$, $SE = 0.18$, $\chi^2\ (1) = 0.0004$, $p = 0.99$; Agr./Dis.: $\beta = –0.22$, $SE = 0.17$, $\chi^2\ (1) = 1.69$, $p = 0.19$).[16]

We ran two additional confirmatory analyses. First, we analysed only accuracy trials for critical meanings in the two conditions. Participants in the Positive Disagreement conditions were predicted to have higher accuracy on these trials relative to the alternative systems using the same type of general particles. Second, we analyzed only preference trials. Participants in the Positive Disagreement conditions were predicted to have a stronger preference for using the specialized particle. For the +/– conditions (two levels, Positive Agreement and Positive Disagreement, sum coded), this analysis revealed an effect of condition on accuracy on critical meanings ($\beta = –3.00$, $SE = 0.65$, $\chi^2(1) = 25.37$, $p < 0.001$). However, this was driven by poorer performance in the Positive Disagreement condition, contrary to our prediction. No effect of condition was found on preference trials ($\beta = 0.71$, $SE = 0.65$, $\chi^2(1) = 1.21$, $p = 0.27$). For the Agr./Dis. conditions (two levels, Negative Agreement and Positive Disagreement, sum coded) there was no effect of condition on accuracy on critical meanings ($\beta = –0.18$, $SE = 0.44$, $\chi^2(1) = 0.18$, $p = 0.67$). However, for preference trials, there was a significant effect of condition ($\beta = 1.83$, $SE = 0.42$, $\chi^2(1) = 21.19$, $p < 0.001$), with participants in the Positive Disagreement condition being more likely to choose the specialized particle than participants in the Negative Agreement condition. This is in line with our predictions.

3.3 Discussion

In Experiment 2, we tested the hypothesis that specialized particles should map to more marked meanings. We operationalized specialized particles as those which have a restricted distribution and can be analysed as conveying a feature combination (e.g., Positive Disagreement) rather than a single feature (e.g., Positive, or Disagreement). This is in line with Claus et al.'s findings (2017) for German and Farkas & Roelofsen's subsequent interpretation of these results (2019).

---

[16] We ran an additional *posthoc* analysis comparing accuracy rates by condition in the initial testing phase, where participants in both conditions are learning the same two general particles. We found no evidence for an effect of condition among either the +/– conditions ($\beta = 0.001$, $SE = 0.10$, $\chi^2(1) = 0.0004$, $p = 0.99$), or the Agr./Dis. conditions ($\beta = –0.05$, $SE = 0.19$, $\chi^2(1) = 0.07$, $p = 0.80$), suggesting that any differences we find between conditions once the specialized particle has been introduced are not likely to be due to differences in learning of the general particles up to that point. We also compared the two pairs of conditions to each other, using general particle meaning as a fixed effect (two levels, Agr./Dis. and +/–, sum coded). We found that participants were better at learning the general particles in the initial training phase in the Agr./Dis. conditions ($\beta = –0.72$, $SE = 0.10$, $\chi^2\ (1) = 53.68$, $p < 0.001$). The mean accuracy in initial training for the two +/– conditions was around 70%, compared to 88% for the Agr./Dis. conditions.

We trained learners on three-way response particle systems in which there were two general particles, each used to convey one feature, and a third specialized particle which was restricted in its distribution and only conveyed a single feature combination. Across four conditions, the specialized particle expressed either Positive Disagreement (as in languages like French and German), Positive Agreement, or Negative Agreement (the latter two both unattested based on the typological sample reported in Farkas & Roelofsen 2019). Under the hypothesis that specialized particles are mapped to more marked meanings, we would predict that the systems in which the specialized particle is mapped to the most marked meaning according to the hierarchy in (1)—Positive Disagreement—should be learned more easily, and specialized particles used more readily for that meaning. The results of the experiment provide some evidence for this hypothesis, and some evidence that, on it's face, goes contrary to this hypothesis.

In support of the hypothesis, when comparing systems in which the general particles convey Agreement and Disagreement, participants were more likely to actually *use* the specialized particle when it conveyed Positive Disagreement compared to when it conveyed Negative Agreement. Participants in the Negative Agreement condition were nearly equally likely to use the general particle available for this meaning in the system. This is very much in the spirit of Roelofsen & Farkas (2015) and Farkas & Roelofsen (2019), who argue that specialized particles should be used to convey marked meanings. More specifically, this result provides evidence in favor of a central (and non-obvious) assumption behind the hierarchy in (1): that polarity mismatch between the response and its antecedent discourse contributes to semantic markedness. As outlined in the Introduction, this is the key feature—not negativity or feature (in)congruence—that Roelofsen & Farkas (2015) argue distinguishes Negative Agreement from Positive Disagreement, the most marked meaning in the hierarchy.

By contrast, when comparing systems in which the general particles were used to convey Positive or Negative polarity, participants were not more likely to use the specialized particle when it conveyed Positive Disagreement than when it conveyed Positive Agreement, despite the fact that Positive Agreement is the *least marked* meaning in the hierarchy. In addition, participants found it *easier* to learn the system with the specialized Positive Agreement particle. This is potentially in line with the results from Experiment 1, which suggested that Positive Agreement forms were generally easiest to learn, and were preferentially mapped to short particles whenever that was possible in the language. However, there is an alternative explanation for the learning difference between these two conditions, which relies on the potential influence of English (i.e., participants' native language) on participants' performance.[17] Recall that in the +/– conditions, the general particle conveying Negative polarity was always the foil in critical accuracy trials (see **Table 5**). In order to convey the critical meaning (either Positive Disagreement or Positive

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

Agreement), participants had to choose either a general Positive particle, meaning something similar to English 'yes', or the new specialised particle, over the general Negative particle, which might be taken to mean something similar to English 'no'. If the general particles were indeed interpreted as English 'yes' and 'no', then participants in the Positive Disagreement condition might have struggled in these accuracy trials, since both 'yes' and 'no' can be used to convey this meaning in English. Thus, the general Negative particle could have been a salient alternative to express Positive Disagreement in all accuracy trials. This would not be an issue for participants in the Positive Agreement condition: only 'yes' can be used to convey Positive Agreement in English, and thus a foil particle similar to 'no' is easier to reject in all cases. Interestingly, **Figure 6** suggests that participants in the Positive Disagreement condition struggle specifically with trials involving the two general particles, *not* with those involving the specialized particle.[18]

To explore this result statistically, in a exploratory (non-preregistered) analysis we analysed accuracy by condition and type of critical accuracy trial in the +/– conditions. Models included random by-participant intercepts and slopes. This analysis revealed no significant effect of trial type ($\beta = 0.29$, $SE = 0.56$, $\chi^2(1) = 0.30$, $p = 0.58$), but a significant interaction between trial type and condition ($\beta = 0.86$, $SE = 0.37$, $\chi^2(1) = 6.12$, $p = 0.013$). As can be seen in **Figure 6**, this is driven by a difference in performance across trial types in the Positive Disagreement condition, which is not found in the Positive Agreement condition: participants in the Positive Disagreement condition were less accurate when they had to choose between two general particles than when the choice involved the specialized particle. This suggests that participants struggled specifically when they encounter trials where (a) both alternative answers could be mapped into English 'yes' and 'no'; and (b) the context allows both 'yes' and 'no' responses in English.[19]

Could this influence of English also make learning harder for participants in *both* of the Agr./Dis. conditions? There are good reasons why this may not be the case. In these conditions, participants learn a system with one general particle conveying Disagreement (potentially more similar to English 'no'), and another one conveying Agreement (potentially more similar to English 'yes'). General accuracy trials involved choosing between these two particles. If participants perceive these to be similar to English 'no' and 'yes' respectively, then they might once again struggle in the Positive Disagreement condition, where both 'yes' and 'no' are possible in English. Crucially, the same is true in the Negative Agreement condition: in English, the

[18] We thank an anonymous reviewer for suggesting this possibility.

[19] It's not entirely obvious why the analogy with the English system would play a role only in trials where participants have to choose between general particles. Given that the English particle 'no' is a possible answer in Positive Disagreement contexts, one might expect participants to also struggle in trials involving the specialized particle. An anonymous reviewer suggests that the influence of participant's native language might be reduced for particle-meaning mappings that are expected under the hierarchy. In other words, perhaps participants do not struggle with the specialized particle exactly because it is conveying Positive Disagreement, a marked meaning according to the hierarchy.

Negative Agreement meaning can also be conveyed by both 'yes' and 'no' particles. However, the general particles in these conditions are arguably less similar to English 'yes' and 'no', as English has a preference for conveying absolute features (Goodhue & Wagner 2018; Repp et al. 2019). The particles here may then not present participants with the same difficulty.[20] Indeed, in a second exploratory analysis, we found no significant main effect of trial type across Agr./Dis conditions ($\beta = 0.22$, $SE = 0.32$, $\chi^2(1) = 0.49$, $p = 0.48$).[21]

To summarize, in this experiment, we find some evidence in line with the predictions derived from Roelofsen & Farkas (2015). In particular, the contrast in behavior between participants in the two Agr./Dis. conditions is consistent with the prediction that the markedness hierarchy influences the likelihood with which specialized particles are mapped to particular meanings during learning. More specifically, our findings support the idea that two contexts, both involving negativity and feature incongruence (Positive Disagreement and Negative Agreement), differ in semantic markedness as predicted by the hierarchy in (1). While we did not find such evidence in the +/– conditions, this may be due to an asymmetry in the degree to which participants' experience with the English particle system disrupted their learning.

Below we report one final experiment in which we attempt to provide some evidence that English speakers are sensitive to *feature incongruence*. This is a key characteristic of the hierarchy proposed by Pope (1976) and Roelofsen & Farkas (2015), for which we have not yet found any behavioral evidence.

4 Experiment 3

In Experiments 1 and 2, we looked for evidence that learners prefer to map more marked particle forms to meanings that are hypothesized to be more marked (e.g., Positive Disagreement, Pope 1976; Roelofsen & Farkas 2015). Across these experiments, we operationalized formal markedness in two different ways: length and restricted distribution. In both cases, we found

[20] Indeed, most participants here reported that the meanings were something like 'agree vs. disagree' or 'true vs. false', rather than 'yes' and 'no'.

[21] This exploratory analysis also revealed an unexpected significant interaction between trial type and condition ($\beta = -0.56$, $SE = 0.19$, $p = 0.002$). As can be seen in Figure 6, participants in the Positive Disagreement condition were better when the general particle was the correct choice compared to when the specialized particle was the correct choice. This is the opposite of what we saw for the Positive Disagreement (+/–) condition. No such difference was found in the Negative Agreement condition. It is not entirely clear what might explain this pattern, and it is worth cautioning again that this was an exploratory analysis. Arguably, it could be related to the fact that participants were generally good at learning the meanings of the general particles in Agr./Dis conditions (see Footnote 16), and potentially less good at learning the specialized particles. Participants in the Positive Disagreement condition may be particularly good at associating this meaning with Disagreement and therefore choosing the right one of the two general particles. In the Negative Agreement condition they may not be quite so good at associating Negative Agreement with the Agreement meaning. This is in line with the idea that there is a natural affinity between Negation and Disagreement (Pope 1976; Roelofsen & Farkas 2015).

some evidence in favour of the hierarchy. In Experiment 1, learners showed sensitivity only to the least marked meaning, Positive Agreement, mapping the shorter form to this meaning across conditions. There was no evidence, for example, that they perceived Positive Disagreement as more marked than Negative Disagreement, since they were not more likely to associate the former meaning with a longer particle. As we suggested above, this points to the possibility that learners could be sensitive to the markedness of both Negative absolute and relative polarity (with respect to Positive Agreement), but not to contrasts due to feature (in)congruence alone. In Experiment 2, we found that meanings that involve disagreement were treated as more marked than those involving negative polarity (i.e. Negative Agreement vs. Positive Disagreement). Both of these meanings are incongruent. No difference was found between Positive Agreement and Positive Disagreement, two meanings that differ both in negativity and feature congruence. Thus, Experiment 2 does not allow us to directly test an effect of feature incongruence only.[22]

In Experiment 3, we aim to provide some indication that English speakers are sensitive to feature (in)congruence. This will help us to determine whether learners' failure to map incongruent meanings to more marked forms in Experiment 1 was because they do not actually perceive these meanings as more marked than their congruent alternatives. Here, we focus on meanings involving disagreement and ask whether contexts involving Positive Disagreement are perceived as *semantically* more marked than contexts involving Negative Disagreement.

One could imagine various ways of cashing out differences in semantic markedness across these contexts. For example, contexts that lead to Positive Disagreement could be more surprising or unexpected than other contexts. As observed in the Introduction, there is no evidence suggesting that Positive Disagreement meanings are substantially less frequent than e.g. Negative Disagreement ones. This suggests that these meanings may not necessarily be particularly unexpected.

Feature incongruence arises specifically in contexts where the antecedent of the response has negative polarity. Differences in semantic markedness could then have their root in comprehension or production cost associated to the computation of negation (truth-reversal operation). In the specific case of disagreement scenarios, processing a Positive Disagreeing response might be cognitively harder than a Negative one because it involves computing two negations rather than just one (i.e., negating the negative antecedent). This is compatible with the results reported by

[22] In principle, we could have tested this by comparing two systems in which the specialized particle conveyed either Positive Disagreement or Negative Disagreement—these two meanings differ only in feature (in)congruence. However, this pair of conditions would suffer the exact same potential confound we identified in the Positive Disagreement vs. Positive Agreement condition of Experiment 2. As with Positive Agreement, Negative Disagreement is expressed in English by a unique particle (in this case 'no'). In contrast, both 'yes' and 'no' can be used to convey Positive Disagreement. We saw in Experiment 2 that this asymmetry in form-to-meaning mapping in participants' native language apparently influenced learning, leading to difficultly interpreting the results. The same would be true of this hypothetical comparison.

Noveck et al. (2021), who find that when French speakers answer a negative question, they are slower to produce the specialized particle 'si' than the underspecified 'oui' and 'non'.

Here, we build on these results, capitalizing on the intuition that scenarios that involve Positive Disagreement are costly, and therefore impact successful communicative interactions between interlocutors. The idea is that, for the speaker, disagreeing with their interlocutor is always costly, but it's even more costly when their interlocutor's utterance involves negation. First, as discussed, disagreeing with a negative statement is arguably cognitively more demanding than other possible responses. Second, negative statements are often less informative than their positive counterparts (Horn 1989). Signaling disagreement with a less informative statement might suggest a stronger disagreement. As a result, these scenarios might be more likely to give rise to an argument between interlocutors.

We showed English-speaking participants interactions involving either agreements or disagreements. We then asked them to rate which of two interactions—one involving a positive polarity response, and the other involving a negative polarity response—would be more likely to cause an argument. If participants indeed perceive Positive Disagreement as more marked than Negative Disagreement, then we predict that they will be more likely to choose interactions involving a positive response in disagreeing contexts. By contrast, in agreement contexts, we do not expect absolute polarity to have an impact on the likelihood of leading to an argument, as both negative and positive responses involve agreement.[23] This experiment, including exclusion criteria and main statistical analyses, was preregistered: https://osf.io/rvy24. We also report additional post-hoc analyses that were not preregistered, as noted below.

4.1 Methods

4.1.1 Stimuli

Response particles were used in the context of a conversation, where two sisters discuss aspects of the life of a family member, as in **Figure 7**. In each trial, the younger sister provides a statement about their family member, and her older sister responds, using a response particle (either 'yes' or 'no') to convey one of the four meanings: Positive Agreement, Negative Agreement, Positive Disagreement and Negative Disagreement. Illustrations for each of these contexts are provided in **Table 6**. The family member was one of {Grandma, Grandpa, Aunt Gia, Uncle Gio}, randomly assigned by participant. Scenarios used for the statement and response concerned purported facts about the family member's life, likes and dislikes, friends and family, etc., discovered by the little sister in a diary. There were 20 unique scenarios used without repetition across the experiment,

[23] We should note, however, that judging which of two agreement scenarios is more likely to lead to an argument is quite unnatural. We return to this issue in the Discussion and thank an anonymous reviewer for raising this point.

as described further below (full list in appendix A.1). Scenarios were balanced with respect to whether they involved discussing typically positive, typically negative, or neutral facts.

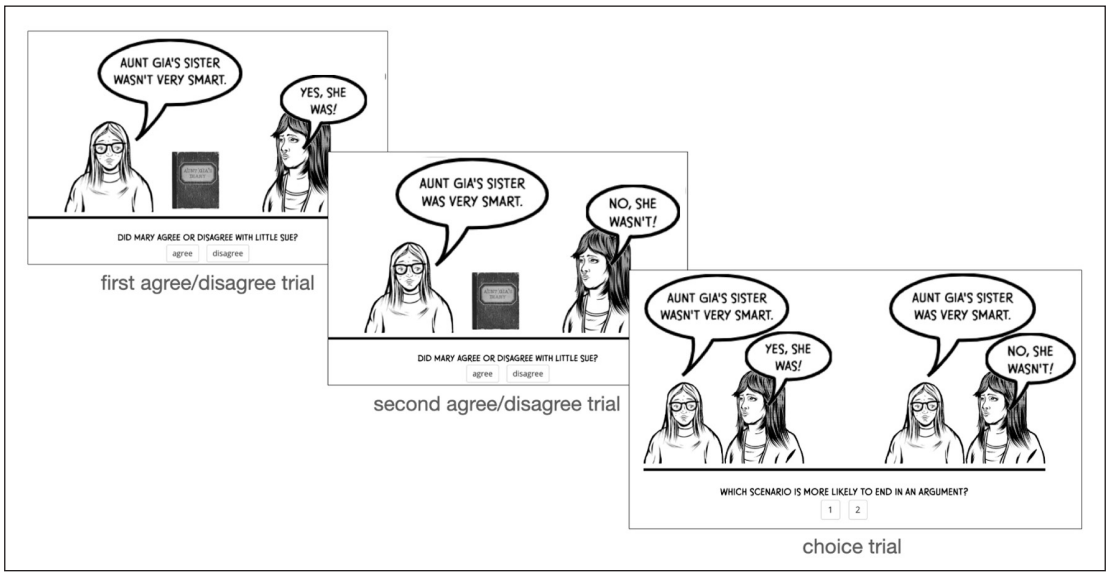

Figure 7: Experiment 3 testing trial procedure (disagree trial). NB: Exposure trials each resemble a single agree/disagree trial.

	Agreement	**Disagreement**
Positive	L.S.: Aunt Gia's sister was very smart. M.: Yes, she was!	L.S.: Aunt Gia's sister wasn't very smart. M.: Yes, she was!
Negative	L.S.: Aunt Gia's sister wasn't very smart. M.: No, she was not!	L.S.: Aunt Gia's sister was very smart. M.: No, she was not!

Table 6 Illustration of meaning contexts in Experiment 3. Different interactions between Little Sue (L.S.) and Mary (M.).

4.1.2 Procedure

Participants were first introduced to two sisters—Mary and Little Sue—from a large family living in New York city, who love each other but sometimes argue. They were told that the little sister had a school assignment to write a report about a family member (Grandma, Grandpa, Aunt Gia, or Uncle Gio) and found an old diary to use. Then participants were told they would be watching interactions between two sisters in which Sue would tell Mary something she put in her report about the family member, but Mary would not always agree.

The experiment consisted of two phases, exposure and testing. In the exposure phase, participants watched an interaction and had to indicate whether Mary agreed or disagreed with

Sue by clicking the corresponding button (see **Figure 7**). Feedback was given. There were 20 trials total, each featuring a randomly chosen scenario, 3 trials for each of the four meaning contexts (Positive Agreement, Negative Agreement, Positive Disagreement and Negative Disagreement) randomly ordered by participant.

During testing, participants were told that they would watch pairs of hypothetical interactions. For each interaction in the pair, participants first watched the interaction and indicated whether Mary agreed or disagreed with Sue, just as before. Feedback was given. Then both interactions were shown on the screen, and participants were asked to choose which interaction was most likely to result in an argument between the sisters (as in **Figure 7**).

Pairs of interactions always featured the same randomly chosen scenario, adjusted for the context. Pairs were either both Agreement contexts, or both Disagreement contexts. In other words, they were matched on whether Mary agreed or disagreed with Sue. Within each type of context, the interactions differed in the polarity of the response—either Positive or Negative. There were 10 trials total, 5 pairs of Agreement contexts, and 5 pairs of Disagreement contexts. Statements were not repeated across pairs and were distinct from scenarios shown to a given participant during exposure.

4.1.3 Participants

All participants were recruited through the Prolific Academic online recruiting platform and gave informed consent. They received 1.50 GBP as compensation. We recruited a total of 50 participants, with the Prolific pre-screening criterion that they be monolingual English speakers.

4.2 Results

In this experiment, we tested whether the semantic markedness associated with feature incongruence would result in participants treating Positive Disagreement scenarios as more likely to lead to an argument than Negative Disagreement ones. The prediction was that participants should consistently select positive over negative polarity responses in Disagreement scenarios. By contrast, this should not be the case in Agreement contexts, where neither response would be particularly likely to lead to an argument (because both responses involve agreement). Consequently, we expect at chance responses in Agreement scenarios (although see Footnote 23).

Following our pre-registered plan, we excluded testing trials in which participants responded incorrectly with ‘agree’ or ‘disagree’ when watching either of the interactions in a pair (8% of trials, mean accuracy 97%). For the remaining trials, participants’ choices were coded as a binary response variable with choice of the Positive response as 1 and choice of the Negative response as 0. Average choice of Positive responses in Agreement and Disagreement trials is shown in **Figure 8**. We analysed this data in two steps.

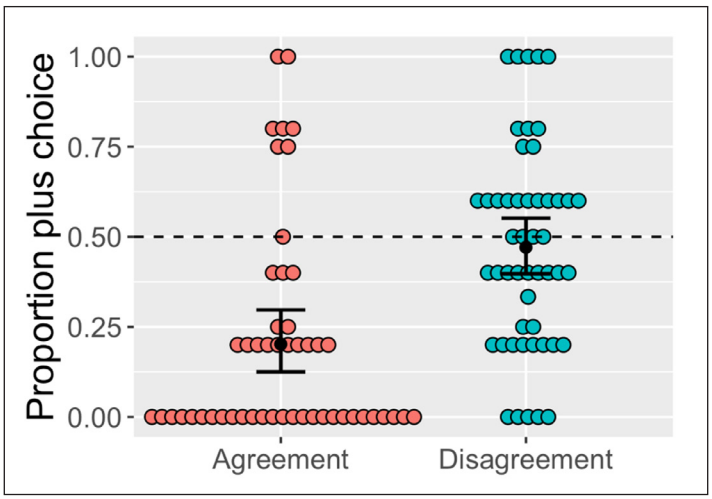

Figure 8: Choice of positive polarity response as more likely to trigger an argument in Agreeing and Disagreeing pairs in Experiment 3. Dots are individual participant's means, error bars are 95% bootstrapped confidence intervals.

First, we tested whether the choice of the Positive response as more likely to trigger an argument was modulated by type of context (Agreement or Disagreement). Qualitatively, the results in **Figure 8** seem to support this. To assess this quantitatively, we fit two mixed-effects logistic regression models. The first model included only an intercept term; the second model included the fixed effect of context type (Agreement vs. Disagreement, sum coded, with Agreement ordered first). Both models included by-participant random intercepts as well as by-participant random slopes for trial type. A Likelihood Ratio Test confirmed that trial type significantly improved model fit ($\chi^2(1) = 46.90$, $p < 0.001$). Inspection of the model revealed a significant intercept indicating an above-chance selection of the negative polarity option, ($\beta = -0.89$, $SE = 0.15$, $p < 0.001$), but an effect of trial type such that positive polarity responses were more likely to be chosen for Disagreement scenarios ($\beta = -0.75$, $SE = 0.12$, $p < 0.001$). In other words, although participants do generally based their choices on the presence of 'no' in the response, when both trials involve a disagreement, they are more likely to choose the Positive Disagreement context as potentially generating disagreement.

As a follow up, we tested the stronger prediction that in Disagreement trials, participants would be more likely to choose the Positive response as triggering an argument than the Negative response. A mixed-effects logistic regression model conducted on the data from Disagreement trials, with only an intercept term, including by-participant random intercepts, revealed no significant difference from chance ($\beta = -0.015$, $SE = 0.17$, $p = 0.38$). In other words, participants are equally likely to choose Positive and Negative response scenarios on Disagreement trials.

A post-hoc analysis further confirms that the interaction between trial type and polarity found in our main analysis was mainly driven by the fact that, in Agreement scenarios, participants

consistently choose Negative responses as more likely to trigger an argument than Positive responses ($\beta = -2.1$, $SE = 0.42$, $p < .001$).

4.3 Discussion

Feature (in)congruence is an important driving force behind the hierarchy in (1). It accounts for the impact of the antecedent to the semantic markedness of a response. Responses conveying incongruent features (i.e., Negative Agreement and Positive Disagreement) are those that involve replying to a negative antecedent. Thus the claim that these responses are more marked equates to the claim that there is some markedness *coming from* the negative antecedent. In this final experiment, we investigated whether English speakers are sensitive to feature (in)congruence by testing whether they consider positive polarity responses to negative antecedents as more likely to lead to an argument than negative responses to positive antecedents.

Contrary to our predictions, participants in our experiment did not treat positive responses as more likely to trigger an argument than negative polarity responses in the context of a disagreement. Assuming that the likelihood of leading to an argument correlates with semantic markedness, participants did not straightforwardly treat Positive Disagreement contexts as more marked than Negative Disagreement.

However, our findings did reveal a relationship between the absolute polarity of the response and whether the response agrees or disagrees with the previous statement (i.e. the relative feature). In agreement contexts, participants consistently treated negative polarity responses as more likely to trigger an argument. In contrast, when participants were faced with disagreement scenarios, this changed, and positive polarity responses were treated as just as likely to trigger an argument.

Participants' responses in Agreement scenarios suggest that the presence of an overt negation in the response and in the antecedent is associated with disagreement in the sense of likelihood to trigger an argument, even when there is no *real* disagreement. This feature of our results might reflect some general bias to treat negative polarity as a potential argument-trigger. This general bias would play a stronger role in Agreement scenarios, where either both antecedent and response involve negation or none does, than in Disagreement contexts, where both scenarios involve some instance of negation. However, an alternative interpretation is that participants develop a specific strategy (i.e., relying on the polarity of the response/antecedent) to deal with Agreement scenarios, and simply choose randomly in Disagreement contexts. This strategy might be prompted by the unnaturalness of the task in agreement contexts. Either way, participants appear to associate negative statements with 'disagreement' regardless of whether the scenario involves agreement or disagreement.

This last possibility was further explored through a post-hoc analysis of how quickly participants indicated whether Mary agreed or disagreed with Little Sue in a given scenario.

Recall that in both exposure and testing trials participants started by watching an interaction between the two sisters and had to indicate whether Mary agreed or disagreed with Little Sue. We analysed how long participants took to decide whether an interaction involving a positive or a negative response involves an agreement or a disagreement. That is, the time recorded from the moment each scenario was presented to participants until response (see **Figure 7** for reference).[24] **Figure 9** shows the average time taken to provide correct responses for each type of interaction. Response times were log-transformed for the statistical analysis. A linear mixed-effects regression model revealed a significant interaction between the type of response (Agreement or Disagreement) and the absolute polarity of the response ($\beta = 0.11$; $SE = 0.018$; $\chi^2(1) = 39.5$; $p < 0.001$): in Agreement trials, participants took significantly longer to respond in negative ($M = 2712$ ms.) than in positive ($M = 1975$ ms.) scenarios ($\beta = 0.121$; $SE = 0.027$; $\chi^2(1) = 17.5$; $p < 0.001$). However, in Disagreement trials, they took longer to respond in scenarios involving a positive answer ($M = 2532$ ms.) compared to a negative one ($M = 1962$ ms.; $\beta = -0.106$; $SE = 0.026$; $\chi^2(1) = 14.65$; $p < 0.001$).

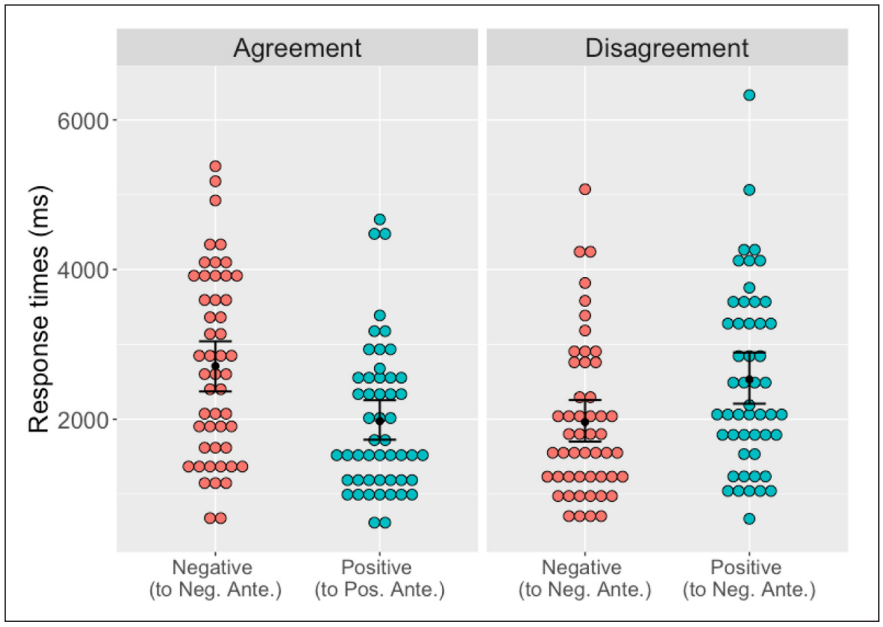

Figure 9: Time (in ms) taken to judge whether a positive or negative statement is in Agreement or in Disagreement with previous discourse (Antecendent). Dots represent individual participant's means, error bars are 95% bootstrapped confidence intervals.

[24] We opted to analyze response times rather than accuracy rates because accuracy was very high across the board for these trials. We analyzed all trials where participants had to decide whether the interaction involved agreement or disagreement (i.e., both exposure and testing trials). We excluded from the analysis trials with response times below 200ms and above 12426 ms., which corresponds to the mean response time across all trials plus one standard deviation. This led to the exclusion of 2.87% of trials.

This post-hoc analysis reveals that participants find negative statements more difficult to categorize as 'agreeing' with previous discourse than positive statements. Interestingly, they also find positive responses to negative antecedents more difficult to categorize as 'disagreeing'. These results suggest that there is some association between disagreement and negative polarity, on the one hand, and agreement and positive response polarity, on the other.

To summarize, the results of Experiment 3 suggest that speakers do not consider Positive Disagreement to be more likely to lead to an argument than Negative Disagreement. However, these two scenarios do differ in how easy it is to categorize each of them as disagreements. Arguably then, Positive Disagreement scenarios are indeed more marked than Negative Disagreement ones along this dimension. However, this difference may not lead learners to make any implicit assumptions about e.g., markedness in terms of length of the form, as was tested in Experiment 1.

5 General Discussion

Here we have reported a series of three experiments investigating the hypothesis that response particle systems across languages are shaped by differences in the markedness of the meanings they convey. More specifically, we sought behavioral evidence for the hierarchy in (1) (repeated in (2) below), proposed by Roelofsen & Farkas (2015); Farkas & Roelofsen (2019), following previous work by Pope (1976) and others. This hierarchy was motivated by the existence of three-way response particle systems in languages like French, German, and Dutch, which have been analyzed as involving a specialized particle conveying Positive Disagreement.

(2) Positive Agreement < Negative Disagreement < Negative Agreement < Positive Disagreement (where '<' indicates less marked).

In our first two experiments, we looked for evidence that this hierarchy might shape typology through learning. In Experiment 1, we tested whether learners would expect forms that are more marked according to (2) to be longer, in line with a general and well-documented trend for more marked meanings to be expressed with more marked forms (e.g., Kurumada & Grimm 2019; Haspelmath 2021). Across four conditions, each of which compared a pair of meanings on the hierarchy, we exposed participants to a response particle system where a longer full form and a shorter clipped form were both used equally frequently for the critical pair of meanings (e.g., *vadof* vs. *va* used for both Positive Agreement and Positive Disagreement). If learners are sensitive to the hierarchy as predicted, we should see them condition this free variation on the markedness of the meaning—the long form should be used more for the more marked meaning in the pair. We found mixed evidence for this. In particular, when Positive Agreement was among the critical meanings, participants preferred to use the short clipped form to express it. However, participants showed no evidence of a preference to use the longer form to express more marked

meanings (e.g., Positive Disagreement, Negative Agreement). Moreover, we failed to find any evidence for an effect of semantic markedness when the pair of meanings did *not* involve Positive Agreement. This suggests that Positive Agreement may have a special status, with other aspects of the hierarchy not exerting any influence on learning at least when it comes to this kind of length-based form markedness.

In Experiment 2, we operationalized markedness in a different way: in terms of restricted distribution. We exposed participants to a three-way particle system which had a restricted particle, used with only one of the meanings in the hierarchy. In one set of conditions, this specialized particle was used either for Positive Disagreement (the most marked meaning according to the hierarchy) or for Positive Agreement (the least marked meaning). In the other set of conditions the specialized particle was used with either Positive Disagreement or Negative Agreement (the second least marked meaning). The systems with a specialized particle for Positive Disagreement are more similar to how Farkas & Roelofsen (2019) analyze the response particles in French and German; the alternative systems are argued to be unattested. If learners are sensitive to the hierarchy in this case, we expect that learning and using a specialized particle for Positive Disagreement will be easier. We found evidence that participants more readily *used* the specialized particle for Positive Disagreement as compared to Negative Agreement. However, participants were particularly good at *learning* in the Positive Agreement condition (potentially because this meaning is in general very easy to learn).

Finally, in Experiment 3, we looked for additional evidence that English speakers are sensitive to feature incongruence—something we failed to find in Experiment 1—by testing whether Positive Disagreement contexts are perceived by English speakers as more likely to trigger an argument between interlocutors compared to Negative Disagreement contexts. We did not find direct evidence for this. However, we found an interaction between whether the response agrees or disagrees with the antecedent and the polarity of the response: negative polarity responses responses are generally perceived to be more likely to lead to an argument, but this is less the case in disagreement cases. In line with this result, we additionally found that participants took longer to determine whether contexts involving feature incongruence (Positive Disagreement and Negative Agreement) involved disagreement vs. agreement.

Taken together, these results provide some evidence for the connection between (three-way) response particle systems and the hierarchy of markedness hypothesised by Roelofsen & Farkas (2015) and others. However, they suggest that different aspects of formal markedness might tap into different aspects of the hierarchy. Learners were sensitive to the special markedness in meaning hypothesized for Positive Disagreement when markedness in form was cashed out in terms of restricted use—the sense of markedness perhaps most close to what Roelofsen & Farkas (2015) have in mind—but not when it was cached out in terms of length. Our findings also

point consistently to a special role for Positive Agreement which is not highlighted by Roelofsen & Farkas (2015) in their predictions about the typology of particle systems. In Experiment 1, participants preferred to use short forms for this meaning, in both Experiments 1 and 2 they found particles mapped to these meanings easiest to learn, and in Experiment 3 they found these contexts easy to identify as involving agreement. These results all point to participants' high level of sensitivity to the *unmarked* status of Positive Agreement. Based on this, we might expect to find not only evidence for systems designed around a special way of marking Positive Disagreement, but also systems designed to specially mark *Positive Agreement*. For example, languages might show a tendency to express this meaning with a particularly short form, or even with a specialized particle restricted to this context (see Holmberg 2016 for a brief mention of such particles in responses to statements). Put more generally, the relationship between markedness an specialized forms might differ from what Roelofsen & Farkas (2015) suggest; for example, perhaps both highly marked *and* highly unmarked meanings can be mapped to specialised forms.

With this in mind, it is worth considering again the analysis proposed for languages like French in Roelofsen & Farkas (2015) and Farkas & Roelofsen (2019). Recall that under this view, particles like 'si' are analysed as expressing Positive Disagreement, a feature combination. By contrast, particles like 'oui' and 'non' are analysed as expressing absolute features (Positive or Negative polarity). But this analysis depends crucially on the claim that these two less restrictive particles have some degree of overlap: for example, 'oui' should be used in *both* Positive Agreement and Positive Disagreement contexts. Given that the evidence for this is not clear for French (see Pasquereau 2020a), it's worth noting that if the 'oui' particle *were* used only in Positive Agreement contexts, then it would have to be analysed as exactly the kind of particle Roelofsen & Farkas (2015) argue does not exist. That is, the system would not necessarily have a single specialized particle, but two. Future work on the typology and usage distributions of response particle systems is needed to determine whether only the most marked meaning can have a restricted distribution, or also the least marked one.

Finally, it is notable that we have not provided evidence here for all aspects of the hierarchy. It remains possible that our experiments are failing to capture some crucial feature of the usage of these particles, or that participants are learning the particles in a way that prevented us from finding the predicted effects in all cases. While we have outlined some possible issues above, it remains for future work to determine in what specific ways and to test these. It is likely to be important to test speakers of languages other then English to ensure that any evidence (or lack thereof) linking learning to the typology of response particle systems is independent of previous language experience.

6 Conclusion

Response particles, like English ‘yes’ and ‘no’, Japanese ‘hai’ and ‘iie’ and French ‘oui’, ‘non’ and ‘si’, are used to respond to polar questions and assertions. These systems are universally found in languages, but languages differ in the number of such particles they have, and in how the particles are mapped to meanings. Theoretical and typological work on response particle systems has argued that these systems are organized in a way that reflects properties of the meanings to be conveyed. Some meanings, like Positive Disagreement—where a response disagrees with a preceding positive question or assertion—have been argued to be especially marked, while others, like Positive Agreement, have been argued to be relatively unmarked. In previous work, Roelofsen & Farkas (2015) and others have hypothesised a hierarchy of meaning markedness, and argued that more marked meanings, like Positive Disagreement, are more likely to conveyed by specialised particles used *only* for these meanings, or by longer forms (e.g., multi-word phrases). Across a series of artificial language learning experiments, we tested whether participants are sensitive to the markedness distinctions posited by these theories. We found consistent evidence that Positive Agreement was treated as special; it was easiest to learn particles mapped to this meaning, participants preferred that it be expressed using a short form, and participants were quick to identify it as an agreeing response. We also found some evidence of sensitivity to the markedness of Positive Disagreement. Participants took longer to determine that Positive Disagreement in fact indicated disagreement, and in some cases preferred to use a specialised particle for expressing Positive Disagreement. However, this aspect of markedness did not always influence the systems participants learned, and other distinctions in the hierarchy were not clearly evidenced in their behavior. Thus while our findings support some aspects of the markedness hierarchy, they also suggest that the hierarchy and the data underlying it may be missing the important role of Positive Agreement in shaping response particle systems.

A Appendix

A.1 Experiment 3 scenarios

The 20 scenarios used for Experiment 3 are listed below. Note that 'X' could be filled in with 'Grandma', 'Grandpa', 'Uncle Gio', or 'Aunt Gia'. The polarity of the statement was determined by the context: positive for Positive Agreement and Negative Disagreement; negative for Negative Agreement and Positive Disagreement.

i. X's favorite holiday was/wasn't Christmas.
ii. X had to/didn't have to do a lot of chores.
iii. X grew up/didn't grow up very poor.
iv. X worked/didn't like working in an office.
v. X's first trip abroad was/wasn't to the US.
vi. X went/didn't go to a expensive school.
vii. X's oldest son was/wasn't +pronoun+ favorite.
viii. X's mom was/wasn't Sicilian.
ix. X's best friend was/wasn't from Italy.
x. X learned/didn't learn to speak English.
xi. X wanted/didn't want to learn to drive.
xii. X's dad had/didn't have a good job.
xiii. X's first job was/wasn't in the family store.
xiv. X had/didn't have a lot of friends.
xv. X liked/didn't like +pronoun+ teachers.
xvi. X's sister was/wasnt' very smart.
xvii. X's favorite thing to cook was/wasn't pasta.
xviii. X visited/didn't visit Silicy more than once.
xix. X wanted to be/didn't want to be a doctor.
xx. X's brother was/wasn't very religious.

Data accessibility statement

All data and analyses are available at https://tinyurl.com/yc6ua7eu.

Ethics and consent

The study was approved by the Ethics Committee of the School of Psychology, Philosophy and Linguistics at the University of Edinburgh. All study participants gave informed consent prior to participation.

Acknowledgements

The authors wish to thank Floris Roelofsen and Anders Holmberg for useful comments and suggestions, as well as to three anonymous *Glossa* reviewers. This research was supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 757643).

Funding information

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 757643).

Competing interests

The authors have no competing interests to declare.

Reference

Barr, Dale J & Levy, Roger & Scheepers, Christoph & Tily, Harry J. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language* 68(3). 255–278. DOI: https://doi.org/10.1016/j.jml.2012.11.001

Bates, Douglas. 2010. lme4: Mixed-effects modeling with R. http://lme4.r-forge.rproject.org/book.

Bellugi, Ursula. 1967. *The acquisition of the system of negation in children's speech*. Harvard Graduate School of Education dissertation.

Blanchette, Frances. 2015. *English negative concord, negative polarity, and double negation*. City University of New York.

Claus, Berry & Meijer, A. Marlijn & Repp, Sophie & Krifka, Manfred. 2017. Puzzling response particles: An experimental study on the German answering system. *Semantics and Pragmatics* 10. 19. DOI: https://doi.org/10.3765/sp.10.19

Culbertson, Jennifer & Franck, Julie & Braquet, Guillaume & Barrera Navarro, Magda & Arnon, Inbal. 2020. A learning bias for word order harmony: evidence from speakers of non-harmonic languages. *Cognition* 204. DOI: https://doi.org/10.1016/j.cognition.2020.104392

Culbertson, Jennifer & Newport, Elissa L. 2015. Harmonic biases in child learners: In support of language universals. *Cognition* 139. 71–82. DOI: https://doi.org/10.1016/j.cognition.2015.02.007

Culbertson, Jennifer & Smolensky, Paul & Legendre, Géraldine. 2012. Learning biases predict a word order universal. *Cognition* 122. 306–329. DOI: https://doi.org/10.1016/j.cognition.2011.10.017

De Swart, Henriette. 2009. *Expression and interpretation of negation: an ot typology*, vol. 77. Springer Science & Business Media.

Diller, Anne-Marie. 1984. *La pragmatique des questions et des réponses*, vol. 243. Tübingen: Gunter Narr Verlag.

Farkas, Donka F. & Roelofsen, Floris. 2019. Polarity particles revisited. *Semantics and Pragmatics* 12(15). 1–16. DOI: https://doi.org/10.3765/sp.12.15

Fedzechkina, Maryia & Jaeger, T. Florian & Newport, Elissa L. 2012. Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences* 109(44). 17897–17902. DOI: https://doi.org/10.1073/pnas.1215776109

Ferdinand, Vanessa & Kirby, Simon & Smith, Kenny. 2019. The cognitive roots of regularization in language. *Cognition* 184. 53–68. DOI: https://doi.org/10.1016/j.cognition.2018.12.002

González-Fuente, Santiago & Tubau, Susagna & Espinal, M. Teresa & Prieto, Pilar. 2015. How do languages reject negative propositions? typological evidence on the use of prosody and gesture. *Frontiers in Psychology* 6(art. 899). 1–17. DOI: https://doi.org/10.3389/fpsyg.2015.00899

Goodhue, Daniel & Wagner, Michael. 2018. Intonation, yes and no. *Glossa: a journal of general linguistics* 3(1). DOI: https://doi.org/10.5334/gjgl.210

Greenberg, Joseph. 1966. *Language universals*. The Hague: Mouton.

Haspelmath, Martin. 2021. Explaining grammatical coding asymmetries: Form–frequency correspondences and predictability. *Journal of Linguistics* 57(3). 605–633. DOI: https://doi.org/10.1017/S0022226720000535

Holmberg, Anders. 2013. The syntax of answers to polar questions in english and swedish. *Lingua* 128. 31–50. DOI: https://doi.org/10.1016/j.lingua.2012.10.018

Holmberg, Anders. 2016. *The syntax of yes and no*. Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780198701859.001.0001

Horn, Laurence R. 1989. *A natural history of negation* (The David Hume Series). Stanford, Calif: CSLI.

Hudson Kam, Carla & Newport, Elissa. 2009. Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology* 59(1). 30–66. DOI: https://doi.org/10.1016/j.cogpsych.2009.01.001

Jones, Bob Morris. 1999. *The welsh answering system*. De Gruyter Mouton. DOI: https://doi.org/10.1515/9783110800593

Kanwal, Jasmeen & Smith, Kenny & Culbertson, Jennifer & Kirby, Simon. 2017a. Language-users choose short words in predictive contexts in an artificial language task. In *Proceedings of the 39th annual meeting of the cognitive science society*. Austin, TX: Cognitive Science Society.

Kanwal, Jasmeen & Smith, Kenny & Culbertson, Jennifer & Kirby, Simon. 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: https://doi.org/10.1016/j.cognition.2017.05.001

Kramer, Ruth & Rawlins, Kyle. 2009. Polarity particles: an ellipsis account. In *Proceedings of nels* 39. 479–92.

Krifka, Manfred. 2013. Response particles as propositional anaphors. In *Proceedings of semantics and linguistic theory* 23. 1–18. DOI: https://doi.org/10.3765/salt.v23i0.2676

Kurumada, Chigusa & Grimm, Scott. 2019. Predictability of meaning in grammatical encoding: Optional plural marking. *Cognition* 191. 103953. DOI: https://doi.org/10.1016/j.cognition.2019.04.022

Leslau, Wolf. 1962. 'yes' and 'no' in the ethiopian languages. *Language* 38(2). 147–148. DOI: https://doi.org/10.2307/410875

Maldonado, Mora & Culbertson, Jennifer. 2020. Person of interest: Experimental investigations into the learnability of person systems. *Linguistic Inquiry*, 1–42. DOI: https://doi.org/10.31234/osf.io/nxbd3

Moser, Elena Vera. 2019. *Polarity-reversing affirmative particles: A feature of standard average european (sae)*: University of Stockholm Ph.D. thesis.

New, Boris & Pallier, Christophe & Brysbaert, Marc & Ferrand, Ludovic. 2004. Lexique 2: A new french lexical database. *Behavior Research Methods, Instruments, & Computers* 36(3). 516–524. DOI: https://doi.org/10.3758/BF03195598

Noveck, Ira & Petit, Nicolas & Tian, Ye & Turco, Giuseppina. 2021. Revealing pragmatic processes through a one-word answer: When the french reply si. *Journal of Memory and Language* 120. 104245. DOI: https://doi.org/10.1016/j.jml.2021.104245

Pasquereau, Jérémy. 2020a. French polar response particles and neg movement. *Natural Language Semantics* 28(4). 255–306. DOI: https://doi.org/10.1007/s11050-020-09164-w

Pasquereau, Jérémy. 2020b. Polar response particles in french as remnants of ellipsis. *Glossa: a journal of general linguistics* 5(1). DOI: https://doi.org/10.5334/gjgl.1064

Pope, Emily. 1976. *Questions and answers in english*: MIT Phd.

R Core Team. 2020. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing Vienna, Austria.

Reali, Florencia & Griffiths, Thomas L. 2009. The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition* 111(3). 317–328. DOI: https://doi.org/10.1016/j.cognition.2009.02.012

Repp, Sophie & Meijer, A. Marlijn & Scherf, Nathalie. 2019. Responding to negative assertions in germanic: On yes and no in english, dutch and swedish. In *Proceedings of sinn und bedeutung*, vol. 23. 267–286.

Roelofsen, Floris & Farkas, Donka F. 2015. Polarity particle responses as a window onto the interpretation of questions and assertions. *Language* 91(2). 359–414. DOI: https://doi.org/10.1353/lan.2015.0017

Sadock, Jerrold M. & Zwicky, Arnold M. 1985. A note onxy languages. *Linguistics and Philosophy* 8(2). 229–236. DOI: https://doi.org/10.1007/BF00632367

Saldana, Carmen & Herce, Borja & Bickel, Balthasar. 2022. A naturalness gradient shapes the learnability and cross-linguistic distribution of morphological paradigms. In *Proceedings of the annual meeting of the cognitive science society*.

Saldana, Carmen & Smith, Kenny & Kirby, Simon & Culbertson, Jennifer. 2021. Is regularisation uniform across linguistic levels? Comparing learning and production of unconditioned probabilistic variation in morphology and word order. *Language Learning and Development*. DOI: https://doi.org/10.1080/15475441.2021.1876697

Smith, Kenny & Wonnacott, Elizabeth. 2010. Eliminating unpredictable variation through iterated learning. *Cognition* 116(3). 44–449. DOI: https://doi.org/10.1016/j.cognition.2010.06.004

Storkel, Holly L. 2013. A corpus of consonant–vowel–consonant real words and nonwords: Comparison of phonotactic probability, neighborhood density, and consonant age of acquisition. *Behavior Research Methods* 45(4). 1159–1167. DOI: https://doi.org/10.3758/s13428-012-0309-7

Takagaki, Yumi. 2014. Oui au lieu de si: les usages exceptionnels de oui/si/non dans les textes écrits. In *4e congrès mondial de linguistique française*, vol. 8. 2917–2932. EDP Sciences. DOI: https://doi.org/10.1051/shsconf/20140801180

Tian, Ye & Breheny, Richard. 2016. Dynamic pragmatic view of negation processing. In *Negation and polarity: Experimental perspectives*, 21–43. Springer. DOI: https://doi.org/10.1007/978-3-319-17464-8_2

Trabasso, Tom & Rollins, Howard & Shaughnessy, Edward. 1971. Storage and verification stages in processing concepts. *Cognitive psychology* 2(3). 239–289. DOI: https://doi.org/10.1016/0010-0285(71)90014-4