Person of interest: Experimental investigations into the learnability of person systems

Mora Maldonado and Jennifer Culbertson

December 13, 2019

1 Introduction

Person systems—typically exemplified by pronominal paradigms (e.g. 'I', 'you', 'she')—serve to categorize entities as a function of their role in the context of speech: there is the speaker, the addressee and others, who play no active role in the conversation. As in other semantic domains, it has long been observed that person systems exhibit what appears to be constrained variation across languages: some person systems are very frequent, while others are very rare or do not occur at all (Cysouw, 2003; Baerman et al., 2005).[1]

Typological regularities of this sort have led linguists to propose universal constraints on possible person systems (Silverstein, 1976; Ingram, 1978; Noyer, 1992; Harley and Ritter, 2002; Harbour, 2016; Bobaljik, 2008; Ackema and Neeleman, 2018). Such constraints are often conceived of (either implicitly or explicitly) as reflecting characteristics of our linguistic capacity which have consequences for learning: specifically, they are assumed to delimit the space of hypotheses entertained by the learner (Chomsky, 1965; but see Piantadosi et al., 2013). However, while person has been extensively investigated from formal and typological perspectives, the link between hypothetized universal constraints on person systems and learnability remains largely unexplored (though see Nevins et al., 2015 for a related artificial learning approach, and Moyer et al., 2015; Hanson et al., 2000; Hanson, 2000; Brown, 1997 for some acquisition studies). In this paper, we introduce an artificial language learning methodology to investigate the existence of universal constraints on person systems. Before providing more details about our methodology, we introduce some

[1] This kind of typological tendencies have been seen in other semantic domains, involving both content and logical words. For example, cross-linguistic regularities have been argued to provide evidence for a universal basis for color categorization, reflecting properties of the human perceptual system (Gärdenfors, 2004; Kay and Regier, 2007; Zaslavsky et al., 2018; Gibson et al., 2017; Steinert-Threlkeld and Szymanik, 2019). Arguments of this sort have also been made to explain the distribution of kinship systems across languages (Kemp and Regier, 2012; Kemp et al., 2018). Relatedly, the study of logical words has also revealed that connectives and quantifiers found in natural languages only cover a very small subset of all possible meanings of these types, indicating the existence of semantic universals (Barwise and Cooper, 1981; Katzir and Singh, 2013; Steinert-Threlkeld and Szymanik, 2019; Piantadosi et al., 2016; Chemla et al., 2019).

additional theoretical background from which we will derive a number of specific predictions regarding the learnability of these systems.

1.1 The Person Space

As mentioned above, there are three conversational roles typically delimited in the person space: the speaker, the addressee and others, who are not active participants in the conversation. We follow standard assumptions and represent them as i, u and o, respectively (e.g., Harbour, 2016). From this ontology, we obtain seven logically possible person categories or 'meta-persons': i, io, iu, iuo, u, uo, o (Sokolovskaja, 1980; Bobaljik, 2008; Sonnaert, 2018). Research on the typological distribution of person systems, however, has found evidence for only four of these being grammaticalized as person categories in languages: first exclusive (i), inclusive (iu), second (u) and third (o). This asymmetry can be directly captured by assuming that the speaker and addressee are unique–there are no forms which express multiple speakers, or multiple addresses—but there can be an undefined number of others (following Harbour, 2016).[2] The meanings expressed by the unattested combinations (io, iuo uo) can then be captured as the interaction between person and number. Each of the four core person categories can be pluralized by adding extra others (see fn. 2; Boas, 1911). Table 1a illustrates these person and number (expressed by the presence or absence of the subscript o) categories. To account for these, and only these person categories, theories of the person space have traditionally posited two primitive binary features: ±speaker and ±addressee (or other equivalent notations; Ingram, 1978; Silverstein, 1976; Noyer, 1992; Bobaljik, 2008). The interaction between these two binary features predicts all and only the four attested person categories, as can be observed in Table 1b.

An example of a language which makes a 4-way person distinction in its pronominal system is Mandarin. Each person category in Table 2a is expressed by a different pronoun, with additional variations depending on whether the referent is singular or plural. This system has 7 forms total, since the inclusive is inherently not-singular (it necessarily refers to both the speaker and addressee), and thus always features plural morphology in Mandarin. Another example is Ilocano, which differs from Mandarin in that it makes a minimal-augmented number distinction rather than singular-plural. In this system there are 8 forms, including two distinct inclusive forms, one minimal ('ta' = speaker and addressee

[2]This assumption is not trivial. As soon as multiple speakers and addressees are allowed in the ontology, each logically available combination of the three entities—i, io, iu, iuo, u, uo, o—should count as a possible person category independent of number. For example, one could conceive of one form that refers to the speaker alone (i), and another form that refers to the speaker plus someone else (io), each with their plural alternative (ii vs. ioo). However, this contrast is never grammaticalized: no language distinguishes between plural expressions referring to multiple speakers/addressees and expressions referring to the speaker/addressee plus others. Indeed, this has been formulated as a typological universal: Pluralities containing participants (speakers or addressees) are never formally distinguished from pluralities containing others. This universal has been extensively discussed in the literature on person systems, but is not directly investigated in this paper. We refer to the reader to the relevant papers (Greenberg, 1988; Cysouw, 2003; Bobaljik, 2008; Wechsler, 2010).

1st	i_o	+speaker -addressee
INCL	iu_o	+speaker +addressee
2nd	u_o	-speaker +addressee
3rd	o_o	-speaker -addressee

(a) Attested categories **(b)** Binary features account

Table 1: Four persons system

only) and the other augmented ('tayo' = speaker, addressee and others).

In many other languages, the meaning space is partitioned such that not all possible person and/or number categories are expressed by distinct forms. Such languages exhibit homophony. For example, one could distinguish *inclusive* languages from *non-inclusive* languages (terminology from Daniel, 2005) like English, where there is homophony between the first and the inclusive persons (in addition to homophony between 2nd singular and plural, see Table 2). Feature-based accounts of person often derive a restricted set of all possible partitions of the person and number space as defined by the presence or absence of homophony among the cells in the space. Such theories only derive homophony patterns by contrast neutralization or underspecification: a distinction that is made available by the grammar might not be active in a specific language (Halle and Marantz, 1994; Harbour, 2008; Harley, 2008; Pertsova, 2011). Specifically, the set of features in Table 1b straightforwardly derives three person homophony patterns based on which contrasts are left underspecified ($\pm$speaker, $\pm$addressee or both). For example, neutralizing the $\pm$addressee feature would generate syncretism between 1st and INCL categories (grouped as [+speaker]), on the one hand, and between 2nd and 3rd (grouped as [-speaker]), on the other. Other feature-based homophony patterns can also be derived from this system by restricting underspecification to specific natural classes. For example, the aforementioned clusivity distinction is lost when two meanings which share the feature +speaker (i and iu) become indistinguishable (e.g. English). That is, the grouping of 1st and INCL categories relies on them belonging to the same natural class [+speaker]. These kinds of feature-based patterns are often referred to as systematic homophony.

However, many different homophony patterns, have been documented both within and across languages (Cysouw, 2003; Zwicky, 1977; Corbett et al., 2002; Baerman et al., 2005; Baerman and Brown, 2013). Not all of these patterns can be described by feature neutralization. In some cases, two or more meanings which don't share any feature are nevertheless expressed by the same form in a given language (Harbour, 2008). This so-called accidental or random homophony is therefore not described in terms of contrast neutralisation, as the targeted meanings do not belong to the same natural class (e.g., defined by the features in Table 1b). Partitions that are not derivable by a theory are often assumed to arise through historical accident, target mainly individual paradigms in a given language, and be marginal typologically (Sauerland and Bobaljik, 2013; Halle and Marantz, 1994; Pertsova, 2011, but

	MIN	AUG	SG	PL	SG	PL
1st	co	mi	wo	women	I	we
INCL	ta	tayo	–	zamen	–	we
2nd	mo	yo	ni	nimen	you	you
3rd	na	da	ta	tamen	she/he/they	they
	(a) Ilocano		**(b)** Mandarin		**(c)** English	

Table 2: Example personal pronoun systems.

see Cysouw, 2003). But whether the typological evidence accords with this prediction is not always clear.[3]

While estimating the frequency distribution over partitions of the person space is complex, a number of theories have recently been developed to make more fine-grained predictions about possible homophony patterns. Following ideas from phonology (Clements, 1985), Harley and Ritter (2002) put forward a universal feature geometry for person directly based on typological data. Their system, illustrated in Figure 1, is based on three privative features. This derives the same set of four person categories as the binary features account in Table 1b but also establishes hierarchical relations between them, making more accurate predictions about the patterns of homophony which can arise (see also Bejar, 2003; McGinnis, 2005; Cowper and Hall, 2009 for similar approaches). For example, this system derives homophony between 1st, INCL and 2nd categories, and therefore predicts this to arise systematically.

In a similar vein, Harbour (2016) posits a theory specifically designed to capture the robust typological generalization in (1), also known as Zwicky's observation.

(1) Languages that do not have a dedicated phonological form for an inclusive person ('you and us') always assimilate it into the first plural person ('us') and never into the second ('you') or third ('them') (Zwicky, 1977).

[3] As it turns out, determining the cross-linguistic frequency of different partitions of the person (and number) space is not straightforward. It crucially depends on the specific assumptions one makes about how to count different systems. Some authors (Cysouw, 2003; Sauerland and Bobaljik, 2013; Baerman et al., 2005; Baerman and Brown, 2013) include individual person-marking paradigms within a single language. For instance, in English, the pronominal paradigm in Table 2c and the verbal agreement system would be counted separately, as they partition the person space differently. Using this metric for counting, the frequency distribution across language is extremely skewed: In Cysouw's data set, of the 4140 possible partitions of an 8-cells person/number space, only 61 are attested (calculated over 265 paradigms). Another possibility is to count based on an abstract notion of person partition, across all the paradigms in a given language. For example, Harbour (2016) proposes a superposition technique: by overlaying all the paradigms in a given language, one can in principle identify which contrasts are grammaticalized in that language at a more general level. By this metric, a distinction is neutralized in a language if and only if it is inactive across all paradigms. In the case of English the inclusive/exclusive distinction would be neutralized since it is not found in any paradigm, while the first/second person contrast would be present since it it is only neutralised in the verbal agreement system. Under this superposition approach, for example, there are 15 possible partitions of a 4-cells person space, but only 5 are attested typologically.

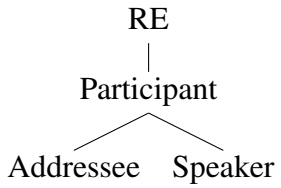


Figure 1: Feature geometry account in Harley and Ritter (2002)

Harbour makes use of two binary features, ±author and ±participant. While the features themselves denote lattices containing referential entities (i, u, iu, o), the values of the features are modelled as complementary operations on lattices. Because features are similar to functions, languages can differ not only in which features are active, but also in the order of feature composition.A similar approach of features as functions has been taken by Ackema and Neeleman (2013, 2018, 2019) with the goal of accounting for the typological tendency in (2).

(2) Languages that feature homophony between first and second plural pronouns ('us' = 'you') and between second and third ('you' = 'them') are far more frequent than those instantiating first-third homophony (Baerman et al., 2005).

Each of these approaches (which we will discuss in more detail below) introduces different theoretical apparatus to capture intriguing typological observations. There are, however, a number of obvious limitations which make basing theories *exclusively* on typological evidence problematic.

1.2 From typology to learning

There is extensive literature now documenting (and in some cases proposing solutions to) the problems posed by typological data samples (for an excellent overview see Cysouw, 2005). For one, such data are generally sparse, and in many cases the number of languages behind a given typological generalisation is quite small. For instance, the largest sample of person/number paradigms, from Cysouw (2003), includes only around 200 languages. Sparse data lead to unreliable estimates of relative frequency, particularly in the tail of the distribution. For example, it is not possible to confidently conclude based on small samples that a given partition is impossible (e.g., see Piantadosi et al., 2013, and also fn 3).

Moreover, typological data are also massively confounded: there are many factors that shape typological distributions (e.g. historical accidents, genetic relations between languages, facts about diachrony; see Dunn et al., 2011; Bickel, 2008; Pagel et al., 2007; Cysouw, 2010, among others), only a subset of which are relevant for building theories of the generative capacity of the linguistic system. The immediate consequence of this is that these data sources cannot be used to argue for a causal link between the cognitive or linguistic system and particular features of language (e.g., see discussion in Culbertson, 2012; Piantadosi and Gibson, 2014; Ladd et al., 2015).

As a response to these general issues—which are relevant for typological data in any domain—there has been an increasing attempt to bring behavioural data on learning to bear on linguistic theories. Specifically, artificial language learning experiments have now been used to link typological universals to human learning and inference in a number of domains including phonology (e.g., Wilson, 2006; Finley and Badecker, 2009; Moreton, 2008; White, 2017; Martin and Peperkamp, sion), syntax (e.g., Goldin-Meadow et al., 2008; Futrell et al., 2015; Tabullo et al., 2012; Culbertson and Adger, 2014; Martin et al., 2019), morphology (e.g., Hupp et al., 2009; Saldana et al., 2019; Fedzechkina et al., 2012) and lexical categorization (Carstensen et al., 2014; Chemla et al., 2019), for reviews see Culbertson (2012); Moreton and Pater (2012); Culbertson (pear).

The present study uses artificial language learning to test a set of predictions derived from the feature-based theories of person described above. By incorporating this new source of data, we can corroborate—or not—the universal constraints on person partitions hypothesized based on typological data. The paper proceeds as follows: in Experiment 1, we establish an experimental set-up to test some basic assumptions of feature-based systems including whether systematic and random homophony are treated differently by learners acquiring a new person system. In Experiment 2, we investigate whether the universal typological tendency known as Zwicky's observation is supported by a learnability advantage, as predicted by Harbour (2016). Finally, Experiment 3 explores potential asymmetries in the learnability of different partitions of 1st, 2nd and 3rd person categories, as predicted by different theories (e.g., Harley and Ritter, 2002; Ackema and Neeleman, 2018).

2 Experiment 1: Something about *us*

Theories of person have hypothesized different inventories of features to constrain the person space and derive only typologically-attested partitions. Regardless of the specific set of features posited, all these approaches assume the features to be *universal*, i.e. part of the human linguistic capacity (Harbour, 2016; Bobaljik, 2008; Harley and Ritter, 2002, among many others). However, not all universal features are necessarily active in a given language: languages can exploit some contrasts and neutralize others (e.g., see English examples in Table 2c). The assumption that a specific set of person features is universal predicts that all things equal humans should have access to, and therefore be able to learn, feature distinctions that are not at play in their native language (L1).[4] Feature-based theories also predict that learners should be sensitive to natural classes: categories that share a feature should be more readily mapped onto the same phonological form. As discussed above, homophony patterns of this type, often called systematic homophony, are privileged by feature-based theories as a natural consequence of an underlying feature-structure. Systematic homophony is predicted to arise regularly and be (easily) learnable. This should

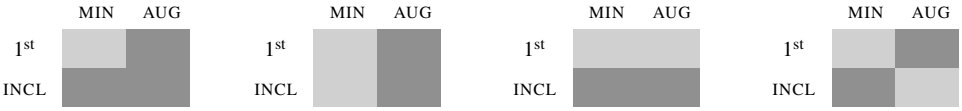
[4]It is of course not always the case that adults retain the ability to make non-native distinctions, for example in phonology (e.g., Werker and Tees, 1984). However, if they cannot in this case, then further experiments will not be possible. Therefore it is necessary to test as a proof-of-concept.

contrast with cases of *random* homophony, where there is no featural basis for two (or more) meanings to share a form.

In Experiment 1, we target these two predictions by focusing on person categories that involve the speaker (first and inclusive persons), and their interaction with number features. For simplicity, we assume the contrasts of interest here arise by the interaction of two binary features: one for person (±addressee) and one for number (±minimal) (see Bobaljik, 2008; Harbour, 2014, for more developed accounts).[5] The resulting 4-cells person space is given in Table 3a. There are multiple partitions of this space defined by homophony. For example, a paradigm that only makes use of the ±minimal number distinction and neutralizes the person contrast (±addressee) would have just two pronominal forms, one for both minimal categories, and another for the two non-minimal (augmented) categories ('Number-contrast' in Table 3c). In contrast, a paradigm that has the 1st/INCL person contrast, based here in the ±addressee distinction, but neutralizes number features would have one inclusive and one 1st exclusive form ('Person-contrast' in Table 3d). Random homophony patterns, not based on feature-based distinctions, are also possible. For example, 1st minimal and inclusive augmented could share the same form, and inclusive minimal and 1st augmented another. The meaning-to-form mapping cannot be expressed here in terms of neutralizing a single semantic distinction ('Random-contrast' in in Table 3e). Put differently, there is no natural class grouping *only* 1st minimal and inclusive augmented. Note that there are three other random partitions of this reduced person space. Let us finally point out that a system like English does not make use of any of these contrasts: English is a non-inclusive language which only makes a number distinction between singular, atomic (e.g. 'I') and plural, non-atomic (e.g. 'we') first person pronouns, not at play here.[6]

	MIN	AUG
1st	(+sp) -add +min	(+sp) -add -min
INCL	(+sp) +add +min	(+sp) +add -min

(a) Reduced person space



(b) English-like **(c)** Number-contrast (Person-homophony) **(d)** Person-contrast (Number-homophony) **(e)** Random

Table 3: Difference partitions of a reduced person space.

[5] We are not committed to this specific feature inventory; our predictions would hold for any theory which posits the contrasts themselves, regardless of the structure of the features space.

[6] Note that there is an asymmetry between the ±minimal and the ±addressee features from the perspective of English: while the ±minimal distinction is completely absent in English, the ±addressee distinction is only non-native *in presence* of the + speaker feature. Put differently, English speakers do in fact have some experience with this feature, which is used to distinguish 1st and 2nd persons. The {+sp, ± add} combination however is non-native.

Here, we investigate whether English-speaking learners are more likely to learn the feature-based patterns in Table 3c-d than the random homophony paradigm in Table 3e. By doing this, we are able to test whether English speakers readily learn contrasts that are not instantiated in their language but are broadly attested typologically, namely the inclusive/exclusive distinction and the minimal/augmented distinction (e.g. in Ilocano, Cysouw, 2003). We can also compare the learnability of systematic and random homophony where neither are present in English. This experiment thus serves as a first step toward a more systematic investigation of universal constraints on person systems; if our predictions are borne out, then we can conclude that the person-number space is indeed based on a set of universal features, such as those posited by the theories described above. We can then test different predictions of particular feature-based theories regarding which specific features are at play. As a sanity check, we also test whether English-speaking learners are biased in favor of a person system that resembles their own, as in Table 3b. If participants perceive the similarity between the new system they are learning and the English person-number system, then this is a good indication that our paradigm is successfully engaging the linguistic space we intend.[7]

To test these predictions, we use two complementary artificial language learning paradigms. In Experiment 1A, participants are taught a pronominal system that matches one of the four paradigms in Table 3, and are then tested on how accurately (within some set number of trials) they are able to learn it ('Ease of learning' paradigm; Culbertson et al., 2017; Tabullo et al., 2012). Accuracy levels are then compared across patterns, revealing which paradigms are easier to learn. Because two paradigms might be equally learnable (within a given number of training trials), but one of them might still be preferred, we also use a second paradigm. In Experiment 1B, we investigate differences in the likelihood of inferring one paradigm or another based on ambiguous data ('Poverty-of-the-Stimulus' paradigm, Wilson, 2006; Culbertson and Adger, 2014). Participants are trained on two cells of the paradigm in Table 3a, and must then use the forms they have learned to express all the cells in the paradigm. In other words, they must extrapolate the taught forms to the remaining two categories. For example, if a learner is trained on two distinct forms for 1st minimal (speaker only) and 1st augmented (speaker plus others), they will be tested on the two remaining categories that include the addressee. If they use the augmented form for both new categories, then they have inferred an English-like paradigm. Different patterns of extrapolation would indicate person or random homophony (as described in detail in Table 4). Experiment 1B would then reveal which paradigms are more naturally inferred by learners in the absence of explicit evidence.

2.1 Methods

Both experiments, including all hypotheses, and predictions, and analyses, were preregistered: Experiment 1A and Experiment 1B. Mterials, data and scripts are provided here. All

[7] We also use a number of other checks to ensure participants' interpret our pronouns as we intend, see below for details.

analyses are as per the preregistration unless we explicitly say otherwise.

2.1.1 Design

Participants in Experiment 1A were randomly assigned to one of four possible conditions: English-like, Number-contrast, Person-contrast and Random-contrast (see Table 3 for reference). Participants in all conditions were taught two pronominal forms mapped into four person categories (1st minimal, inclusive minimal, 1st augmented and inclusive augmented). All conditions instantiate bipartitions of the person space with two-to-one mappings, but differ on which contrast was kept active in the pronominal system (and which one was neutralized): an English-like contrast, a person contrast ($\pm$ addressee), a number contrast ($\pm$ minimal) or a random homophony pattern.

In Experiment 1B, participants were randomly assigned to one of three conditions, illustrated in Table 4.[8] Conditions differed in which subset of two first person categories was trained (critical training set) and held-out (critical held-out set). This determines which alternative full paradigms are consistent with the two categories participants have learned. Condition 1 is consistent with an English-like pattern (or feature-base homophony)[9]. Conditions 2 and 3 are each consistent with one type of feature-based homophony, and random homophony.

All participants in both experiments 1A and 1B were additionally exposed to another four pronominal forms which mapped into the second and third person singular and plural categories. These forms were used as fillers, and were not analysed.

2.1.2 Materials

The same materials were used in Experiment 1A and 1B. In both cases, the language consisted of 6 different pronoun forms, used for the filler categories (2nd sg/pl, 3rd sg/pl), plus the critical first person forms. For each participant, these 6 lexical items were randomly drawn from a list of 8 CVC non-words created following English phonotactics: 'kip', 'dool', 'heg', 'rib', 'bub', 'veek', 'tosh', 'lom'. Items were presented orthographically.

To express the pronoun meanings, we commissioned a cartoonist to draw scenarios involving a family of three sisters and their parents. Each family member has a clearly-defined role in the conversational context. The two older sisters are speech act participants (in all scenarios they are either speaker or addressee). The third (little) sister was spatially close, but never a speech act participant. The parents were seated in the background (serving as additional others).

Pronouns were used as one-word answers to questions like 'Who will be rich?'. Meanings were expressed by visually highlighting subsets of family-members, as in Table 5. In

[8]Two additional conditions were also run, but are not reported here. These were used to test differences between person- and number-based homophony. Since this is orthogonal to the main aim of the experiment, the reader is referred here for details.

[9]It's important to note that the training in Condition 1 is compatible with both a MIN/AUG or a SG/PL system.

	Mappings			Training	Held-out	Compatible paradigms
		MIN	AUG			
Condition 1	1st	A	B	1^{st}_{MIN}, 1^{st}_{AUG}	$INCL_{MIN}$, $INCL_{AUG}$	English-like, Number-contrast, Random× 2
	INCL	A or B?	A or B?			
		MIN	AUG			
Condition 2	1st	A or B?	A	1^{st}_{AUG}, $INCL_{AUG}$	1^{st}_{MIN}, $INCL_{MIN}$	Person-contrast, Random× 3
	INCL	A or B?	B			
		MIN	AUG			
Condition 3	1st	A or B?	A or B?	$INCL_{MIN}$, $INCL_{AUG}$	1^{st}_{MIN}, 1^{st}_{AUG}	Number-contrast, Random× 3
	INCL	A	B			

Table 4: Summary of Conditions in Experiment 1B. There are two training and two held-out categories per condition. Each training category is mapped into a different pronominal form (here called A or B), schematically represented with colours (green/red). Participants can use the training forms they learned (A or B) to express the held-out meanings (gray cells). There are four different paradigms compatible with the training per condition, as specified in the right-most column.

some cases, more than one pattern of visual highlighting could match the target meaning, options were then randomly selected. An example illustrating the $INCL_{.min}$ trial is provided in Figure 2. All questions were English interrogative sentences of the form 'Who will...?', which were randomly drawn from a list of 60 different tokens.

2.1.3 Procedure

The basic procedure was the same in Experiments 1A and 1B. Participants were first introduced to the family, including the names of the sisters, and were told they were going to see the sisters playing with a hat that had two magical properties: whoever wore it could see the future but would also talk in a mysterious ancestral language. Participants were instructed to figure out the meanings of words in this new language. They were given a hint that the words were not names, and an example trial with an English pronoun ('her'). In addition, the speaker and addressee roles switched several times during the experiment to highlight that the words were dependent on contextually-determined speech-act roles. This was induced by swapping who had the magical hat.

Experiment 1A had two phases. In the training phase, participants were first exposed to 6 pronominal forms in the new language, corresponding to the 4 filler and 4 critical person categories. Each exposure trial had two parts: a scene where a question is asked, and a scene where the question is answered with a pronoun form in the language (e.g., Figure 2a). To check that participants were paying attention, they were then asked to select the pronominal form they had just seen from two alternatives. There were 12 training trials (2 repetitions per form). After this initial exposure, participants were given a test of the

Category	Highlighted set
1^{st}_{MIN}	speaker
$INCL_{MIN}$	speaker, addressee
1^{st}_{AUG}	speaker, other(s)
$INCL_{AUG}$	speaker, addressee, other(s)
2^{nd}_{MIN}	addressee
2^{nd}_{AUG}	addressee, other(s)
3^{rd}_{MIN}	one other
3^{rd}_{AUG}	multiple others

Table 5: Highlighted family members for each person category. To ensure that forms were not associated with specific quantities, critical augmented categories randomly include one or two additional others. Third person singular meanings were always expressed with a female other.

trained forms in what we call *what if...* trials. *What if* trials consisted of a question and answer scene, as in exposure, followed by a 'what if?' scene in which a new set of individuals was highlighted. Participants were asked to pick the correct word for that meaning from two alternatives (e.g.,n Figure 2b). There were 32 such trials (3 repetitions per control form, 6 per critical form). Participants were given feedback on their answers. Participants were then given a final critical test. Trials consisted of a question scene, followed by a scene highlighting the referent(s), but no pronominal form. They had to pick the word corresponding to the meaning from two alternatives (e.g., Figure 2c). This phase consisted of 24 trials (3 repetitions per form). Participants received no feedback during this phase.

In Experiment 1B, there was also a training and a testing phase. Crucially, during the training phase, participants were only trained on the pronouns in the filler and critical training sets (6 person categories). There were 12 exposure trials (2 repetitions per form) and 16 *what if...* trials (2 repetitions per filler form, 4 per critical training form). Participants were given feedback on their answers. The critical testing phase included trials for the two remaining critical categories, i.e. the held-out set. This phase consisted of 48 trials (6 repetitions per form). Participants received no feedback during this phase.

Both experiments included a *pre-training* phase where participants were exposed only to the three singular person pronouns. This was done to familiarize participants with the set-up by using the less complex stimuli in terms of highlighting. At the end of both experiments, participants were given a debrief questionnaire, which included questions targeting how they interpreted the meanings they were taught. Importantly, most participants reported having understood the words as pronouns. For example, participants in Experiment 1B (Condition 2) described the meaning of form 'A' as 'Me or us not including you' and the meaning of form 'B' as 'Us including you'.[10] More details about the procedure of these

[10]Not all participants reported pronouns for these meanings. Interpreting participants' responses in these cases is not straightforward. For example, a highly accurate participant reported the meaning of form 'A' to be 'sisters'. Given that this meaning is not consistent with their training, the interpretation becomes difficult.

experiments can be found in Table 10 in the Appendix.

2.1.4 Participants

A total of 197 English-speaking adults[11] were recruited via Amazon Mechanical Turk for Experiment 1A (English-like group: 48, Number-contrast: 50, Person-contrast: 49, Random: 50). 171 participants (English-like group: 44, Number-contrast: 41, Person-contrast: 41, Random: 45) responded accurately on more than 80% of exposure trials during the training phase and were considered for the analyses, according to our preregistered plan. Participants were paid 2 USD for their participation which lasted approximately 15 mins.

A different group of 253 English speakers were tested in Experiment 1B (Conditions 1: 87; Condition 2: 87; Condition 3: 79). Per our preregistered plan, participants were excluded if (a) their accuracy rates during exposure training were below 80%, or (b) they had not answered correctly more than 2/3 of the training trials. This resulted in analysis of 131 participants (Conditions 1: 46; Condition 2: 49; Condition 3: 36).[12] Participants were paid 2.5 USD for their participation which lasted approximately 20 mins.

2.2 Results

2.2.1 Experiment 1A

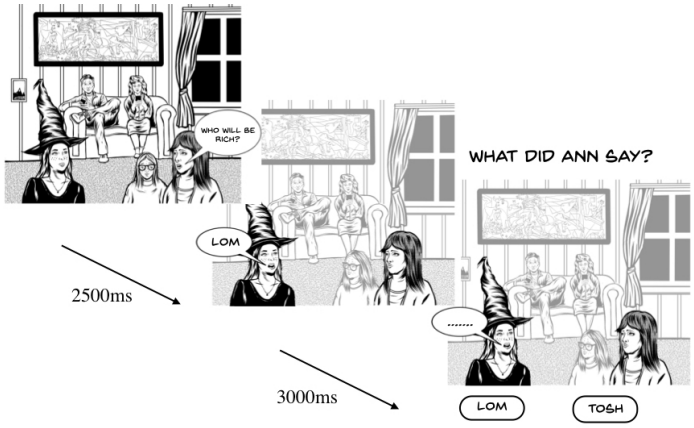
Figure 3 shows the proportion of correct responses in critical trials per experimental condition (English-like, Person-contrast, Number-contrast and Random-contrast) during the testing phase. Per our preregistration, we ran two standard logistic mixed-effects models to evaluate the effect of the experimental condition on accuracy rates (coded as 0 or 1). Both models included random intercepts per subject. The analyses were carried out in the R programming language and environment (R, 2013), using the lme4 software package (Bates et al., 2014). The standard alpha level of 0.05 was used to determine significance, and p-values were obtained based on asymptotic Wald tests.

A first model assessed whether English-like patterns are learned better than alternative paradigms. We used treatment contrast coding with English-like paradigm as the baseline, thus each of the remaining conditions was compared to this fixed level. The output of this model revealed that the proportion of correct responses in the English-Like group was significantly higher than chance ($\beta = 1.78, p < .001$). While responses in the Person-contrast and Number-contrast groups were not statistically different than the baseline (Number-contrast: $\beta = -0.38, p = .28$; Person-contrast: $\beta = -.6; p = 0.093$), accuracy rates in the Random group were significantly lower than the baseline ($\beta = -1.24; p < .001$). This matches the visual pattern in Figure 3.

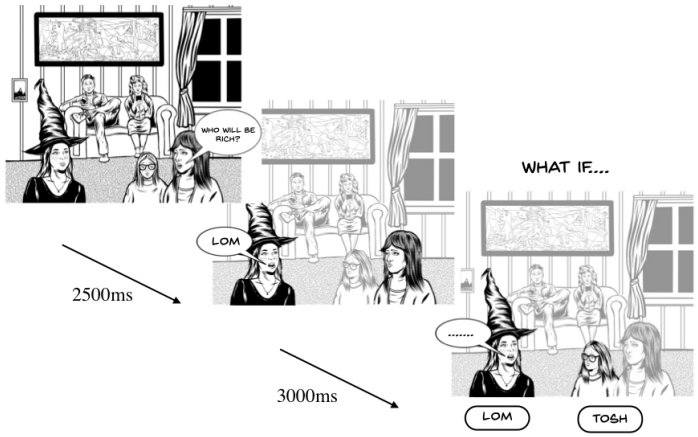
We opted for relying only on accuracy rates to draw conclusions about participants' performance in the experiment.

[11] Two participants were excluded for not being self-reported native speakers of English.

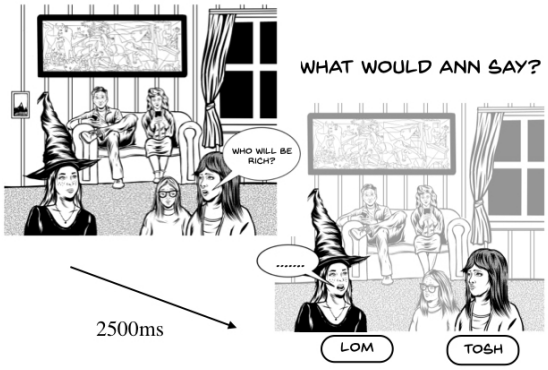
[12] High accuracy rates on trained critical items were required because extrapolation of these forms is not interpretable if participants have not learned them.



(a) Exposure (INCL_{min})



(b) *What if...*($1^{\text{nd}}_{\text{aug}}$)



(c) Testing (INCL_{min})

Figure 2: Illustration of Exposure, *What if...* and Testing trials. Feedback was presented for 2000ms after response Exposure and *What if...* trials.

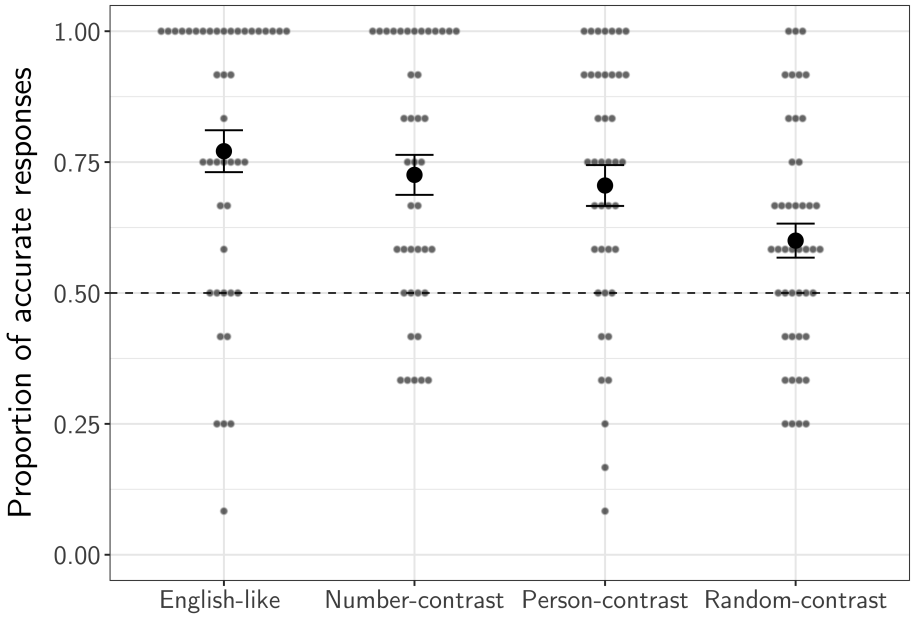


Figure 3: Accuracy rates in *critical* testing trials by condition in Experiment 1A. Error bars represent standard error on by-participant means; gray dots represent individual participants means.s

We ran a second model to explore the difference between feature-based and random patterns. The analysis was restricted to Number-contrast, Person-contrast and Random conditions. We used treatment coding with the Random condition as baseline. The proportion of correct responses in this baseline group was significantly higher than chance ($\beta = .58, p < .001$), but significantly lower than both feature-based conditions (Number-contrast: $\beta = 0.8, p < .01$; Person-contrast: $\beta = 0.62, p = .03$). This suggests that participants trained to make a (non-native) person or number contrast were more accurate than those trained on a random homophony pattern.

2.2.2 Experiment 1B

Recall than participants in Experiment 1B were taught two pronominal forms (coded as forms A and B), which they had to use to describe both a critical trained set, and a held-out set of person meanings involving the speaker (levels: 1^{st}_{min}, 1^{st}_{aug}, INCL_{min}, INCL_{aug}). Figure 4 shows the proportion of trials on which participants chose the form 'A' (pronoun) for each critical person category during the test phase. Choice of the same form across categories indicates homophony. A visual inspection of Figure 4 suggests that participants in Condition 1 are consistently using one form for the 1^{st}_{min} category, and the other for the remaining three categories: this indicates inference of an English-like paradigm. Participants in Conditions 2 and 3 appear somewhat noisier in their responses, however, distinct patterns are evident. In Condition 2, one form is used for the two first person categories,

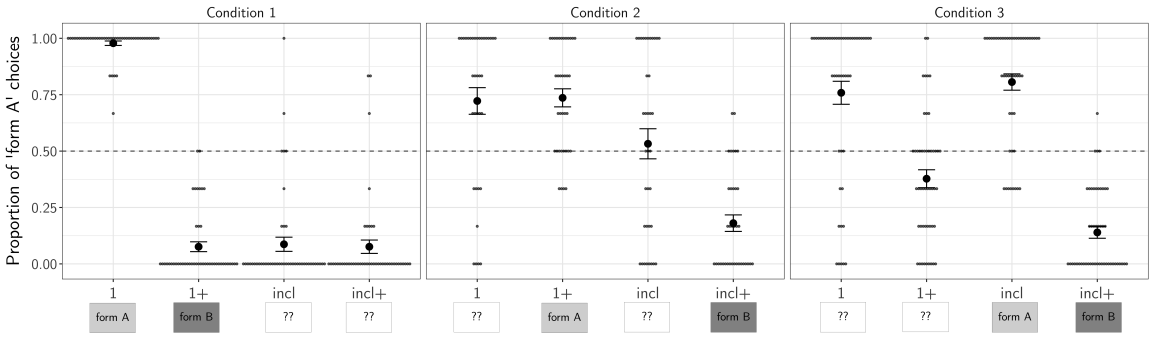


Figure 4: Proportion of form 'A' (as opposed to form 'B') choices for each first person category during the test phase in Experiment 1B. Choice of the same form (A or B) across categories indicates homophony. Error bars represent standard error on by-participant means; dots are means of individual participants.

and, at least for some participants, the other form is used for the two inclusive categories (consistent with maintenance of the person contrast, and number homophony). In Condition 3, one form is used for the minimal categories, and the other for the plurals (consistent with maintenance of the number contrast and person homophony). Note, however, that this figure shows by-participant averages for each category. This does not show clearly which patterns *individual participants* produced.

Figure 4 suggests that there is relatively little variation across participants in Condition 1 compared to the other conditions; almost all participants chose the same form for each category, and they tended to do so categorically. To confirm this statistically, we calculated the *joint entropy* of the held-out set for each individual. This value indicates the degree of uncertainty or variability in each participants' mapping of the taught forms for the two held-out categories. Participants who are less consistent in their answers will have higher joint entropy values. We then fit a simple linear regression model predicting joint entropy by Condition (3 levels). We used treatment coding, with Condition 1 as baseline. No random effects were included in the model, as each participant had a single joint entropy value. As predicted, joint entropy rates were significantly higher for Conditions 2 and 3 (intercept: $\beta = .28$; vs. 2: $\beta = .44 \pm .13,\ p < .001$; vs. 4: $\beta = .64 \pm .12, p < .001$).

A second analysis evaluated whether individual participants in Conditions 2 and 3 were more likely to infer feature-based rather than random patterns (as suggested by Figure 4).[13] We calculated the probability that participants were deriving a feature-based pattern given their held-out answers (see Figure 10 in Appendix for probability of the feature-based pattern per subject and per condition). Depending on the condition, this feature-based pattern could correspond to either a Person-contrast paradigm (Condition 2) or a Number-contrast paradigm (Condition 3). In Condition 2, we computed the probability of choosing form A

[13]Note that this analysis diverges from our preregistered plan, which was designed to address this same question with a different analysis method. We believe the current analysis is both simpler and also more technically sound.

for the $1st_{min}$ and form B for the $INCL_{min}$; in Condition 3, the probability of choosing form 'A' for the $1st_{min}$ and form 'B' for the $1st_{aug}$.

We then ran non-parametric Wilcoxon signed rank tests per Condition to determine whether the probability of deriving a feature-based pattern was higher than chance. Given that there were four paradigms compatible with the training in each condition, chance level was set at 25%. The results of these tests indicate that the probability of deriving a Person contrast paradigm in Condition 2 was not significantly different than chance ($p = .406$), but the probability of deriving a Number contrast paradigm in Condition 3 was above chance ($p < .001$). The same procedure was followed in Condition 1 with respect to the probability of deriving an English-like pattern. As expected, this probability was significantly higher than chance ($p < .001$).

2.3 Discussion

The main aim of these first two experiments was to test whether learners are sensitive to feature combinations, or contrasts, which are not present in their native language. We exposed English-speaking learners to paradigms expressing four person/number categories in a new language. We focused on systems instantiating either the inclusive/exclusive or the minimal/augmented contrasts, which have been argued to have a universal basis in two features, encoding person and number respectively (e.g. ±addressee, ±minimal). We predicted that participants would find paradigms with homophony patterns resulting in these two non-native contrasts more natural than paradigms with random homophony. Our experiments also included an English-like paradigm, as a sanity check. This native-like system was predicted to be preferred over any other.

In Experiment 1A, we tested these predictions by training participants on one of four paradigms, and comparing their ability to learn them. Results confirmed the predicted advantage for native-like pronouns systems: learners found it easier to learn a paradigm with the same structure as English. Results also revealed that participants trained on a pronominal system with a (non-native) person or number homophony pattern outperformed those trained on a random homophony pattern. This supports the claim that learners perceive this 4-cell person space as the interaction of two distinct features, rather than as a conjunction of four different categories, fully independent from each other (in line with Sauerland and Bobaljik, 2013). If learners divide the 1st, inclusive, minimal and augmented categories into two natural semantic classes, one for person and one for number, then one non-native contrast must be learned, but that contrast is based on a natural class.

These results were for the most part confirmed in Experiment 1B, where participants were trained on ambiguous data, and were then required to extrapolate trained forms to new meanings. Learners were more likely to infer an English-like pattern in the face of ambiguous input. They were also more likely to infer person-based than random homophony (making productive use of ±minimal distinction). Results were less clear when number-based homophony was pitted against random homophony; participants were not significantly more likely to infer a paradigms characterized by a person contrast. In other

words, after being trained on an inclusive/exclusive distinction in part of the paradigm (1st and INCL augmented), participants did not generalize this contrast to the rest of the paradigm (the minimal categories). This suggests that although the non-native clusivity distinction is indeed learnable (in Experiment 1A), it may be more difficult for English learners than the non-native minimal/augmented distinction. The apparent difference between number and person homophony is supported by a posthoc analysis showing that accuracy rates on trained categories (before exclusion) are higher in Condition 2 than in Condition 3 ($p < .001$), suggesting that the person distinction was harder to learn than the $\pm$minimal distinction.

One possible explanation for this difference is that it reflects participants' experience with homophony in English. Assuming that English encodes an atomic/non-atomic number distinction, it is possible to characterize English as a case of (only) person homophony (i.e., a non-inclusive language Harbour, 2016). In other words, English speakers have more experience with distinctions in number than in clusivisty. Alternatively, the fact that English speakers do not generalize the person distinction could be thought as the result of a applying a native constraint against such inclusive systems. In contrast with binary features accounts, some approaches (e.g., (Harley and Ritter, 2002)) describe non-inclusive systems as making use of two privative features (Speaker and Addressee) together with a constraint that prevents these features co-occurring. Arguably, the apparent difficulty in generalizing the inclusive/exclusive distinction might suggest that English speakers have a harder time learning a distinction that violates a native constraint against the simultaneous specification of the Speaker and Addressee features.[14]

To summarize, this study presents the first experimental evidence for differences in learnability between alternative person paradigms. Native-like paradigms, unsurpisingly, are easiest to learn and most likely to be inferred when input are ambiguous. More interestingly, paradigms exhibiting homophony within a natural class are learned (and in some cases inferred) more readily than paradigms with random homophony. In what follows, we build on these basic results to investigate more specific constraints on the person space hypothesized by feature-based theories of person.

3 Experiment 2: Within you, without you

In a classic paper from 1977, Zwicky made the following observation regarding the cross-linguistic distribution of person systems: In languages that do not distinguish clusivity (e.g. English), the *you and us* inclusive meaning is always expressed as a form of 'us', and never as a form of 'you' (or 'them').[15]

[14]Note that the advantage for person over number homophony found in Experiment 1B contrasts with typological data, which suggests that languages with different pronominal forms for exclusive and inclusive persons but no number distinction are more common than minimal-augmented languages that do not make an inclusive contrast (Cysouw, 2013). However, these counts are very sparse

[15]This is a generalization about languages and not about individual paradigms within a language, which might show accidental homophony (see Harbour, 2016, and examples therein).

At first glance, Zwicky's generalization, already presented in (1), is quite surprising. Most feature-based approaches to person systems (e.g., Bobaljik, 2008) assume that the inclusive person shares features with both the first exclusive (e.g. +speaker) and the second exclusive (e.g. +addressee). Indeed, a number of languages have inclusive pronounshttps://www.overleaf.com/project/5da086dd95432c0001bb2266 that can be morphologically decomposed into first plus second forms (e.g., in Bislama, Harbour, 2011).[16] This leads naturally to the expectation that languages should be as likely to assimilate the inclusive with the second person as they are to assimilate it with the first. In contrast, no theory would predict the inclusive meaning to be homophonous with the third person, as the inclusive and the third person do not have any features in common (although see Rodrigues, 1990, for a potential exception)).

To account for Zwicky's generalization, there have been two general approaches in the literature. The first maintains the traditional set of features, but posits default feature specifications in order to predict an asymmetry between first-inclusive and second-inclusive (Harley and Ritter, 2002; McGinnis, 2005). The second involves using a different set of features. Harley and Ritter (2002)'s feature-geometry account maintains both Speaker and Addressee features as dependent nodes of the feature Participant, but the Speaker feature is considered to be less marked than the Addressee feature. Consequently, in languages without an inclusive distinction, a preference for assimilating the inclusive meaning into the first person is expected, as they share the default feature. Defaults can be overridden, therefore the second-inclusive homophony pattern can still arise. By contrast, a third-inclusive system is still impossible.

Harbour (2016) posits a different set of features, ±author and ±participant, that denote the semi-lattices $\{i\}$ and $\{i, iu, u\}$ respectively, whereas the values of the features are modelled as complementary operations on lattices. The absence of an ±addressee feature—and of a lattice consisting on only $\{u\}$—creates an inherent asymmetry between speaker and addressee discourse roles.[17] This asymmetry is essential to derive Zwicky's observation as a *strong* constraint on possible person systems. For example, a language which makes use of the ±author feature will have a bipartition of the person space in which i and iu are homophonous and u and o are homophonous. Similarly, a language with both ±author and ±participant can have a tripartition (if ±participant feature composes last, see Harbour, 2016, for details) in which i and iu are homophonous (with two additional forms for u and o), or a quadripartition (if the participant feature composes first). Without a corresponding ±addressee feature, though, there is no way to have any partition which picks out the set including iu and u. Indeed homophony between inclusive and second person or inclusive and third person are both equally impossible.[18]

[16] Although note that there are also languages where the inclusive form patterns morphologically with the second person (e.g. Ojibwa pronouns, Harley and Ritter, 2002 from Schwartz and Dunnigan, 1986).

[17] Harbour's proposed ontology consists of *egocentrically* nested subsets, such that the smallest subset in the ontology contains the speaker alone—i.e. $i \subset i, iu, u \subset i, iu, u, o$.

[18] An intermediate proposal can be found in Ackema and Neeleman (2013, 2018). In their proposed feature structure, "there is no natural class (...) that comprises the first person inclusive and the second person, but not the first person exclusive" (2013, p.910). However, second-inclusive patterns can still be obtained by

The theories outlined above account for Zwicky's observation, but differ critically in how second-inclusive and third-inclusive are treated. Under Harley and Ritter's system (Harley and Ritter, 2002), third-inclusive is singled out as underivable, while second-inclusive is possible but more marked than first-inclusive. By contrast, the system proposed by Harbour (Harbour, 2016) takes as its starting point the idea that only first-inclusive can be generated by the grammar. Based on the typology alone, it is impossible to adjudicate between these theories: both second- and third-inclusive patterns are unattested. Moreover, neither theory provides an explicit mechanism for linking the feature-based representations (and operations) they posit to typology. The implicit link is *learnability*: only a subset of possible person partitions are learnable by humans, or alternatively, some are learned readily while others are more difficult (e.g., can be learned but require substantially more evidence). In Experiment 2, we investigate learners' sensitivity to predicted asymmetries among non-inclusive paradigms. To do this, we use an ease of learning design: we train English-speaking learners on a new language with an inclusive that is a form of 'us' (first-inclusive), a form of 'you all'(second-inclusive) or a form of 'them'(third-inclusive), and compare how well they are learned.

Given that English features first-inclusive homophony *and* this is the only tripartition systematically attested in the typology, learners are predicted to prefer such paradigms over alternatives. As in Experiment 1, this will serve as a sanity check. Regarding second-inclusive and third-inclusive homophony —both unattested in the typology as systematic patterns—the accounts outlined above differ in their predictions. If both of these patterns are directly ruled out by the grammar (*à la* Harbour (2016)), learners should be equally unlikely to learn either of them. By contrast, if learners are sensitive to the semantic commonalities between the inclusive and the second person (e.g. +addressee), a second-inclusive system should be easier to learn than a third-inclusive one (Harley and Ritter, 2002; McGinnis, 2005). This pattern of results would moreover suggest that an apparent asymmetry between first-inclusive and second-inclusive languages should not be encoded as a hard constraint on person systems (contra Harbour, 2016).

3.1 Methods

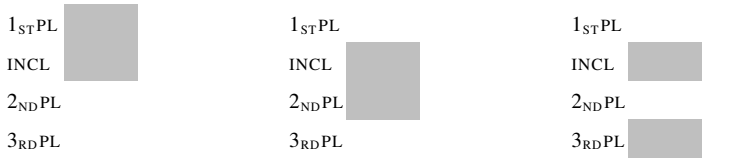
This experiment, including all hypotheses, predictions, and analyses, was preregistered here All materials, data and scripts are provided here.

3.1.1 Design

Participants were randomly assigned to one of three possible conditions, summarized in Table 6. Participants in all conditions were taught three pronominal forms mapped into four *plural* person categories (first exclusive, inclusive, second exclusive, and third). Note that in this experiment we rely on a singular/plural number contrast, not on a minimal/augmented;

incorporating an impoverishment rule in the system (see Footnote 33 in General Discussion for details). This is not possible for inclusive-third homophony, creating an asymmetry between the two unattested patterns.

there is no number distinction within the INCL category. Each condition instantiated a different form-to-meaning mapping: the pronominal system could assimilate the inclusive meaning into the first plural person (First-Inclusive condition), into the second plural person (Second-Inclusive condition), or into the third plural person (Third-Inclusive condition).



(a) First-Inclusive **(b)** Second-Inclusive **(c)** Third-Inclusive

Table 6: Conditions in Experiment 2. Grayed cells are mapping to a single pronominal form, white cells to different and distinct forms.

Participants in all three conditions were moreover exposed to three additional distinct pronominal forms corresponding to the first, second, and third *singular* persons. Participants' learning of these forms was used as an exclusion criteria (see below).

3.1.2 Materials

The language consisted of 6 different pronominal forms: 3 forms were used for the plural pronouns (critical categories), and 3 different forms were used for the singular pronouns (filler categories). For each participant, these 6 lexical items were randomly drawn from a list of 8 CVC non-words used in Experiment 1.

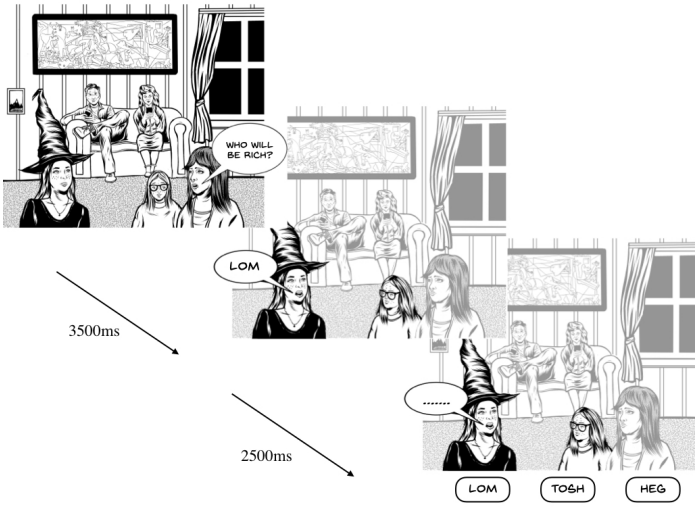
Visual stimuli were the same as in Experiment 1. Pronouns were used as one-word answers to English interrogative sentences of the form 'Who will...?', randomly drawn from a list of 60 different tokens. Meanings were expressed by highlighting a subset of family-member, as in Table 7.

3.1.3 Procedure

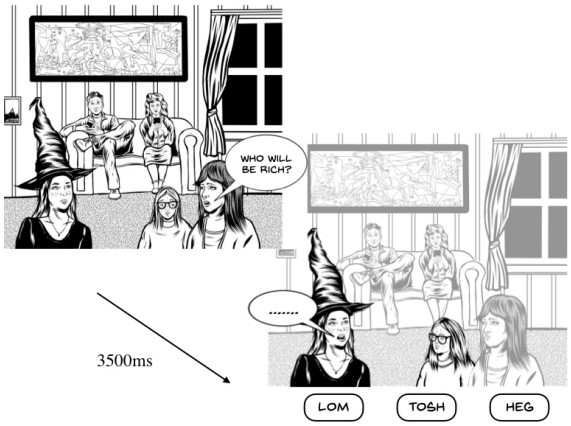
The general backstory was the same as in Experiment 1. Participants were instructed to figure out the meanings of the words in the new language, and they were told that the words they were learning were pronouns. As in Experiment 1, the speaker and addressee roles switched several times during the experiment to highlight that the context dependent pronouns.

The experiment had two phases, each composed by exposure and testing blocks (e.g., Figure 5). Trials in each of these blocks were analogous to the ones used in Experiment 1, except that participants had to select the correct word for that meaning among *three* different options (not two).

During the first phase, participants were exposed and tested on the three singular pronouns. There were a total of 12 exposure and 12 testing trials (4 repetitions per form/meaning).



(a) Exposure



(b) Testing

Figure 5: Illustration of Exposure and Testing trials for the 1stPL category.

Category	Highlighted set
1^{st}_{SG}	speaker
2^{nd}_{SG}	addressee
3^{rd}_{SG}	one other
1^{st}_{PL}	speaker, other(s)
INCL	speaker, addressee (other(s))
2^{nd}_{PL}	addressee, other(s)
3^{rd}_{PL}	multiple others

Table 7: Highlighted family members for each person category. 1^{st}, 2^{nd} and 3^{rd} plural categories randomly include one or two additional others; the inclusive category could refer to speaker and addressee alone or include as well one or two others.

Participants who responded accurately to at least 2/3 of the testing trials in this phase (corresponding to 8 correct responses) passed to the second critical phase. In the critical phase, participants were exposed to and tested on the mapping between three plural pronouns and four person meanings. This phase was comprised of two alternating exposure and testing blocks. There were a total of 24 exposure trials (6 repetitions per meaning) and 48 testing trials (12 repetitions per meaning). More details about the experimental procedure are provided in Table 11.

The complete experimental session lasted approximately 20 minutes. The order of presentation of meanings was fully randomized within exposure and testing blocks for each participant. As in Experiment 1, participants were given a debrief at the end of the experiment to check how they interpreted the forms they were trained on. For example, participants in the Second-Inclusive condition would describe the critical form as 'me and you or you all' or as 'group containing Ann or Mary'.

3.1.4 Participants

A total of 320 English-speaking adults were recruited via Amazon Mechanical Turk (First-inclusive group: 109, Second-inclusive group: 101, Third-inclusive: 110).[19] 167 participants responded accurately on more than 8 singular testing trials and were allowed to continue with the *critical* plural pronoun phase, according to our preregistered plan (First-inclusive group: 57, Second-inclusive group: 55, Third-inclusive: 55). Participants who passed the singular pronoun testing phase were paid 3.5 USD for their participation and 1 USD otherwise.

[19]The number of participants reported here does not include workers who were excluded for not being self-reported native speakers of English (10) and participants who failed to pass an attention check (AC) included at the very beginning of the experiment (35). This AC was added because our exit questionnaires in Experiment 1 revealed that many participants had not read the instruction or were simply bots (Rouse, 2015). While these participants were usually excluded by our other criteria, the AC allowed us to filter them out in advance, distinguishing them from participants who just found the experiment hard.

3.2 Results

Mean accuracy rates on testing trials during the *critical* phase are given in Figure 6. The effect of Condition and Block on accuracy rates was analyzed using logit mixed-effect models in R (R, 2013).[20] The model included the maximal random effect structure, random intercepts per subject and slopes per block (following Barr et al., 2013). The standard alpha level of 0.05 was used to determine significance, and p-values were obtained based on asymptotic Wald tests.

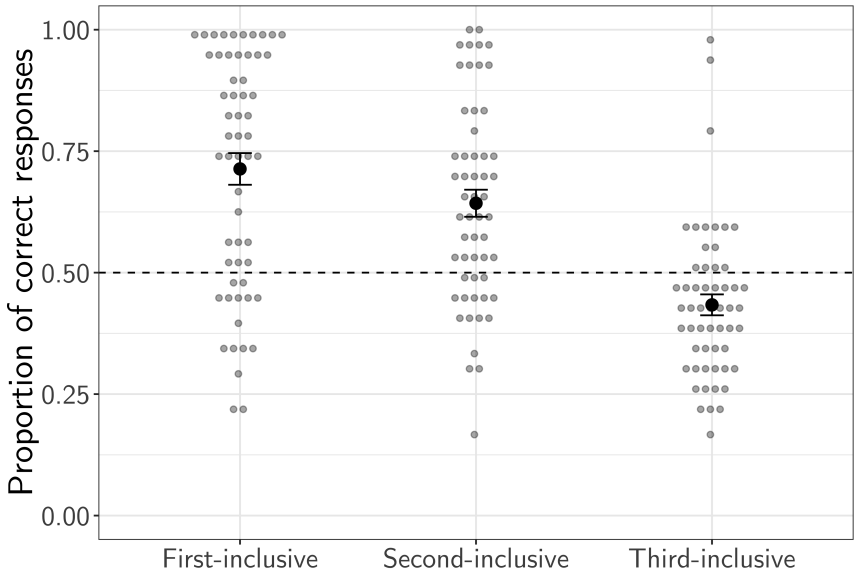


Figure 6: Accuracy rates in *critical* testing trials by condition in Experiment 2. Error bars represent standard error on by-participant means; gray dots represent individual participant means.

We first compared First-Inclusive with Second-Inclusive and Third-Inclusive (contrasts were treatment coded, with First-Inclusive and Block 2 as baselines). The model intercept was significant, indicating that in the First-Inclusive (Block 2) performance was significantly above chance ($\beta = 1.59$; $p < .001$). In addition, accuracy rates in the First-inclusive condition were significantly higher than in the Third-inclusive condition ($\beta = -1.83$; $p < .001$), and marginally different from accuracy rates in the Second-inclusive condition ($\beta = -0.572$; $p = .055$).

A second model was fitted to test the difference in accuracy between Second-inclusive and Third-inclusive conditions (contrasts were treatment coded, with Second-inclusive and

[20]There were two testing blocks in the critical phase, each preceded by an exposure block. Participants were generally expected to improve with accumulated exposure, but this improvement could vary across conditions. Each model included the effect of Block on accuracy, as well as the interaction with Condition. However, we report here only simple effects regarding the second testing block. The complete model output can be found here.

Block 2 as baseline). Accuracy rates in the Second-inclusive condition (Block 2) were found to be significantly above chance ($\beta = 0.97; p < .001$) and higher than the ones in the Third-inclusive condition ($\beta = -1.2; p < .001$).[21]

3.3 Discussion

Our findings first confirm that English-speaking participants are most successful at learning a new language that features native-like homophony between inclusive and first person meanings. As in Experiment 1, this result acts as a sanity check: participants understand the task and are able to learn a language that has the same structure as their own. If, as predicted by Harbour (2016), there is additionally a hard constraint on tripartitions which derives only first-inclusive homophony, then we might expect this preference to be very strong indeed. However, the difference in accuracy between the First-inclusive and Second-inclusive conditions was only marginally significant. This result is consistent with the idea that learners are sensitive to the semantic overlap between inclusive and second person, as predicted by Bobaljik (2008); Harley and Ritter (2002); McGinnis (2005). In line with these accounts, participants in the Second-Inclusive condition might treat these as a natural class, relying on a shared feature, ±addressee, to learn the partition. This result supports a theory in which first- and second-inclusive patterns are both generated by the grammar, but the latter is just dispreferred (contra Harbour, 2016, and possibly Ackema and Neeleman, 2013, 2018, see Footnote 18).[22]

Importantly, we also found that learners have a bias against systems that assimilate the inclusive into the third person. This result reveals that second- and third-inclusive systems, despite being (generally) unattested in the typology, are not equal from a learnability perspective: there is a stronger pressure against third-inclusive than against second-inclusive homophony. This is again as predicted by theories like Bobaljik, 2008; Harley and Ritter, 2002; McGinnis, 2005 (but arguably also by Ackema and Neeleman, 2013, 2018) which posit that, unlike first and second person, third person does not form a natural class with the inclusive category, as it does not have any feature in common with it (cf. e.g., Rodrigues, 1990). As a result, homophony between inclusive and third persons is not predicted to occur systematically, though it might arise accidentally. The learnability cost for third-inclusive systems can be therefore seen as another case of a bias against random homophony.

Returning to Zwicky's observation, our findings suggest that the typological asymme-

[21] Following our preregistration, we ran a second version of each of these models, restricting the analysis to inclusive meanings. The idea behind this move was that our hypotheses target specifically the inclusive category (expressed by the ambiguous pronoun in each system). The output of these models follows the pattern of results described above.

[22] Alternatively, as Daniel Harbour (p.c.) points out, one could argue that our participants are treating the pronominal system we teach them as an instance of accidental syncretism within an inclusive language (i.e., a language that makes an inclusive/exclusive distinction). If this were the case, theories like Harbour (2016) could still account for our results. However, given that our participants are speakers of a non-inclusive language, this seems very unlikely. Further, this would likely not predict the difference we find between second- and third-exclusive.

try between alternative non-inclusive systems may not have a *simple* correlate with learning. In the typology, first-inclusive systems are attested systematically, while second- and third-inclusive systems are not; in our experiment, third-inclusive systems were clearly dispreferred, but the advantage for first-inclusive over second was much weaker. While this is consistent with a theory positing weak learning biases (i.e., constraints that can be over-ridden given sufficient evidence) which penalize third-inclusive most, it still leaves the typological data partially unexplained.

One possibility is that there is an additional weak bias, not at play in our experiments, which further advantages first-inclusive systems. An obvious such candidate is a general *egocentric* bias, i.e., increased importance or salience of the speaker to him or herself (Charney, 1980; Loveland, 1984; Moyer et al., 2015). If individuals perceive the world as a function of their presence in it, they may be more likely to adopt categorization systems which preserve this distinction. This would lead to an asymmetry between first-inclusive and second-inclusive systems. This bias may be weakened in the context of our experiment, where participants are passive learners and do not themselves feature in the meanings they are learning. We return to this issue in the General Discussion.

In the next section, we investigate a second typological generalization that appears to challenge the categorical distinction between participants and non-participants suggested by Zwicky's generalization.

4 Experiment 3: *I me mine*

It has long been observed that there is a fundamental difference between person categories involving speech act participants (first, inclusive and second persons) and the third person, which refers to other, non-participant individuals. Besides having a *fixed* reference, which does not depend on discourse roles, in a number of languages the third person is treated as morphologically distinct from other persons categories (Forchheimer, 1953; see summary in Harley and Ritter, 2002). These facts have often led researchers to propose that the third person is *unmarked* with respect to other persons (Benveniste, 1971). For instance, in Harley and Ritter's (henceforth, H&R) feature-geometry approach, this intuition is captured by treating the third person as the default interpretation of the base node (what the authors call Referring Expression), whereas first, inclusive and second persons require the presence of the dependent node Participant.[23] An illustration of H&R's geometry is reproduced from above in Figure 7.

Despite accounting for a number of interesting typological patterns, H&R's system (and others alike, McGinnis, 2005; Bejar, 2003) is potentially challenged by the cross-linguistic distribution of homophony in person systems. Specifically, these approaches cannot account for the following typological observation, repeated from above [24]:

[23]The Participant node is interpreted by default as first person, and the node in turn dependent on that (Addressee) is interpreted as second person when is the only one specified. The inclusive person is obtained when both Speaker and Addressee nodes are specified together.

[24]This generalisation does not directly concern the inclusive category, as it mostly holds for non-inclusive

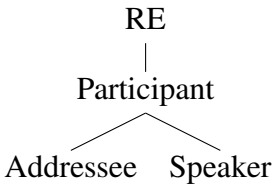


Figure 7: Feature geometry account in Harley and Ritter (2002)

(2) Languages that feature homophony between second and third person categories and between first and second are more frequent than those instantiating first-third homophony (Baerman et al., 2005).

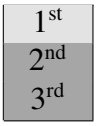
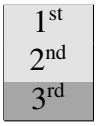
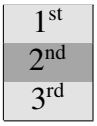
		
2+3 homophony	1+2 homophony	1+3 homophony
Pr: 14/200	Pr: 5/200	Pr: 1/200
Agr: 13/111	Agr: 11/111	Agr: 5/111

Table 8: Illustration of 1+2, 2+3, and 1+3 homophony patterns. Colours indicate forms: cells with the same color use the same form. Typological counts for pronominal systems (Pr) are from Cysouw's data set (Baerman et al. 2005, from Cysouw 2003), and counts for verbal agreement (Agr) are from the Surrey Person Syncretism database Baerman et al. (2005); Corbett et al. (2002).

The numerically higher frequency of first with second (1+2) and second with third (2+3) systems relative to first with third (1+3) systems is illustrated in Table 8. This tendency is found in both free pronouns and in verbal agreement, but it is restricted to non-singular or number-neutral contexts (i.e., it doesn't hold for singular cases).[25] Note that 2+3 homophony is also numerically more common than 1+2 patterns. We return to this in the Discussion section.

Under Harley and Ritter (2002) account, 1+2 homophony can arise through neutralization of the Addressee node, and 1+2+3 can arise through complete underspecification of the Participant distinction. By contrast, there is no way for first-third or 2+3 homophony patterns to be derived via neutralization as these person categories do not share any specific feature. This leads to the expectation that 1+2 homophony will arise systematically and will therefore be more common than both 1+3 and 2+3 homophony, which should in turn

languages, where the inclusive is collapsed with the first person. For the sake of simplicity, we therefore remove the inclusive from the discussion.

[25]The counts inBaerman et al. (2005) for personal pronouns on the one hand, and verbal agreement are drawn from two different typological samples, making the counts not fully comparable to each other. In what follows, we specifically focus on pronoun systems.

be equally unlikely to arise (i.e. only through accidental homophony). This does not match up with the typological counts in Table 8.

Ackema and Neeleman (2013, 2018) propose an alternative theory of person which better accounts for the relative frequency of these homophony patterns. [26].They redefine the person space in terms of two privative person features, dubbed PROX (for "proximate") and DIST (for "distal"). In line with Harbour (2016), they interpret these features as functions that operate on an input set to deliver a subset as output. The crucial aspect of this account for our purposes is that the semantic specification of these two features implies that one is shared by first and second person (PROX), while the other is shared by second and third person (DIST).[27] The immediate consequence is that both 1+2 and 2+3 homophony can be generated, depending on which one of the two features (PROX or DIST) is left underspecified. In contrast, no feature is shared uniquely by the first and the third person (while excluding second person), ruling out the existence of a 1+3 homophony pattern. A similar asymmetry can derived by the theory proposed in Harbour (2016), although he does not specifically discuss it. To summarize, the cross-linguistic observation outlined above is well accounted for by Ackema and Neeleman (2013, 2018) (and Harbour, 2016), whereas it is problematic for Harley and Ritter (2002). However, the typological data this generalization is build on is extremely sparse and the magnitude of the differences is small: of approximately 200 languages in Cysouw's sample (2003), the most frequent 2+3 homophony pattern is attested in the pronominal systems of a mere 14 languages, and the least common 1+3 is attested in one (see Baerman 2005, p.60). The limitations discussed above for typological data are therefore present here in spades. But the issue is an important one: is the traditional and perhaps more intuitive distinction between discourse participants and non-participants central to the organisation of the person space? Or alternatively, is it one asymmetry among potentially many? In what follows, we use the experimental paradigm developed above to bring learnability data to bear on this question.

Using the extrapolation paradigm (cf. Experiment 1B), we measure learners' likelihood of inferring each of the relevant pattern in the face of ambiguous data. Participants are taught the meaning of two pronominal forms, which correspond to a subset of person categories, and they are then tested on how they extrapolate these forms to the remaining category. For example, some learners are taught two distinct forms for 1st and 2nd person, and then tested on which of those forms they use to express the held-out 3rd person meaning. If they use the 1st person form to express the new 3rd person meaning, then they have inferred a 1+3 homophony pattern. A different pattern of extrapolation would indicate 2+3

[26]It's worth noticing that in more traditional accounts of person systems, 2nd and 3rd are also often group as a natural class under the feature [-speaker] (Forchheimer, 1953; Pertsova, 2011)

[27]In a nutshell, Ackema and Neeleman assume that PROX and DIST features operate on a set containing all the possible sets of person referents in the form of nested subsets. That is: the set containing all potential referents (S_{iuo}) contains a subset containing only speaker and addressee referents (S_{iu}), which in turn contains a subset of only speaker referents (S_i). The feature PROX operates on an input set and discards its outermost "layer", whereas DIST selects this outermost layer. It is easy to see that in order to obtain, for example, the first person referent one would need to apply the PROX feature twice, as PROX(S_{iuo}) = (S_{iu}) and PROX(S_{iu}) = S_i.

homophony, as described in Table 9. Harley and Ritter (2002) predict that learners will be more likely to infer 1+2 homophony relative to both 2+3 and 1+3 homophony patterns. Ackema and Neeleman (2013, 2018) predict that learners will be equally likely to infer either 2+3 or 1+2 homophony patterns, but less likely to infer 1+3 homophony.

4.1 Methods

Experimental predictions and analysis pipeline were preregistered here. All materials, data and scripts can be found here.

4.1.1 Design

Participants were randomly assigned to one of three conditions, summarized in Table 9. Conditions differ on which subset of two person categories was used for training and which category was held-out. The training set determines which patterns of homophony participants can extrapolate to: Condition 1 is consistent with 1+2 and 1+3, Condition 2 with 1+2 and 2+3, and Condition 3 with 2+3 and 1+3 patterns. The specific predictions of each account given this design are summarized in the two right-most columns of Table 9. Note that both accounts make the same predictions for Condition 1, but differ in their predictions for Conditions 2 and 3.

All participants were additionally exposed to another two pronominal forms which correspond to the singular alternatives of the plural pronouns they are trained on. The person categories for singular forms were always the same as the critical plural forms for a given participant, and were determined by condition (see table 9). For example, participants in Condition 1 were additionally exposed to 2nd and 3rd *singular* pronouns.

4.1.2 Materials

The language consisted of four different pronominal forms: two plural forms (critical categories) and two singular forms (filler categories). These four lexical items were drawn from the same list of 8 CVC non-words used in Experiments 1 and 2. Visual stimuli were the same as in Experiments 1 and 2. The reference of the pronouns was expressed by highlighting a subset of family-members, as in Table 5, except that in this experiment the inclusive category was never expressed.

4.1.3 Procedure

After being introduced to the general backstory (cf. Experiments 1 and 2), participants were instructed to figure out the meanings of the words in the new language. Participants were given an example trial with an English pronoun ('her' or 'me' depending on the condition) that would help them understand that the words they were learning were pronouns. As in the previous experiments, the speaker and addressee roles switched during the experiment to reinforce the pronominal meanings of the forms.

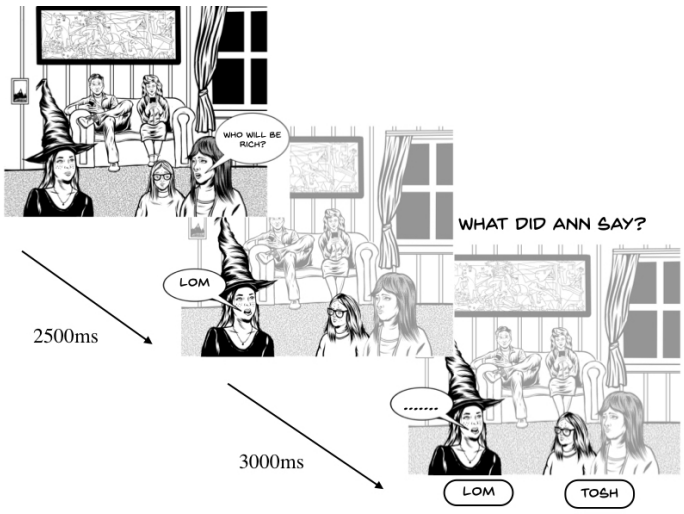
	Mapping		Training	Held-out	Compatible paradigms	H&R	A&N
Condition 1	**1st**	A or B?	2nd, 3rd	1st	1+2, 1+3	1+2 > 1+3	1+2 > 1+3
	2nd	A					
	3rd	B					
Condition 2	1st	B	1st, 3rd	2nd	1+2, 2+3	1+2 > 2+3	1+2 ≈ 2+3
	2nd	A or B?					
	3rd	A					
Condition 3	1st	B	1st, 2nd	3rd	1+3, 2+3	2+3 ≈ 1+3	2+3 > 1+3
	2nd	A					
	3rd	A or B?					

Table 9: Summary of Conditions in Experiment 3. There are two training and one held-out (in bold) categories per condition. Each training category is mapped into a different pronominal form (A or B), schematically represented with shades of grey. Participants must use the training forms they learned (A or B) to express the held-out meaning. The two right-most columns state which of the compatible paradigms participants are predicted to infer under Harley and Ritter (2002, H&R) and Ackema and Neeleman(2013,2018, A&N). accounts.

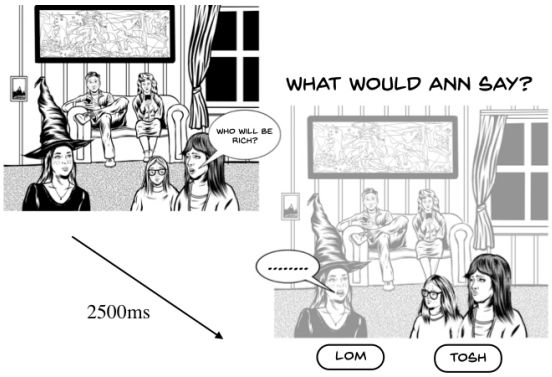
The experiment had two training phases followed by one testing phase, the structure of which were exactly as described for Experiment 1B. The only difference was in the person categories instantiated by the highlighting (see Figure 8). Briefly, the two training phases were composed of exposure and *what if...* trials; the testing phase involved trials in which a referent set was highlighted and participants had to choose the corresponding form. Participants were given feedback after exposure and *what if...* trials, but not after testing trials. The order of presentation of meanings was fully randomized within phases for each participant.

In the first training phase (16 trials), participants were trained on two singular (filler) forms. After this first training, participants were asked to type in a meaning for the two words they had learned, and they were given feedback on their answers.[28] In the second training phase (28 trials), participants were trained on both filler and critical meanings. Finally, in the testing phase the critical held-out meanings were added. There were 24 trials in this phase, eight of which were repetitions of the held-out meanings. As in previous experiments, participants completed a debrief questionnaire at the end of the experiment. A summary of the procedure is given in the Appendix (Table 12).

[28]Unlike Experiment 2, participants were not excluded based on their performance with these singular items in this phase, since participants were told what they mean and they are not used for extrapolation.



(a) Exposure 1stPL category



(b) Testing 2ndPL category

Figure 8: Illustration of Exposure and Testing trials in Experiment 3.

4.1.4 Participants

259 English-speaking adults (Condition 1:74, Condition 2:86, Condition 3:99) who had not participated in one of our previous experiments were recruited via Amazon Mechanical Turk.[29] Per pre-established exclusion criteria (as in Experiment 1), participants who failed to perform accurately in at least 66% of each training category (4/6) during the testing phase were excluded from the analyses. The data of 152 participants were kept for the analyses (Condition 1:48, Condition 2:54, Condition 3:50). All participants were paid 2.5 USD for their participation which lasted approximately 15 minutes.

4.2 Results

Recall that participants were taught two pronominal forms (coded as forms A and B), which they had to use to describe both a critical set of two trained categories as well as a third held-out meaning. Figure 9 shows the proportion of trials on which participants chose the pronominal form coded as 'A' during the testing phase, for each category and condition. For Condition 1, Figure 9 shows a mixed pattern of responses for the held-out meaning (1st plural): some participants used the trained 2nd person form (coded as 'A'), some the 3rd person form (coded as 'B'), and others behaved randomly. By contrast, participants in Conditions 2 and 3 appear to have inferred a consistent paradigm: in both cases, participants largely used the same form (coded as 'A') for 2nd and 3rd person meanings (regardless of what was the trained meaning).

Following our preregistered plan, we ran a logit mixed-effect model predicting form 'A' choices in held-out trials by Condition. Recall that the meaning of the pronominal form coded as 'A' differed depending on Condition (see Mapping in Table 9). We used treatment coding, with Condition 2 as baseline. This baseline ensures that the predictions made by each of the two hypotheses under consideration are distinguishable.[30] The model also included random by-participant intercepts. As before, analyses used the lme4 package in R R (2013) and p-values were obtained based on Type II Wald tests.

[29] Five participants were previously excluded for not being self-reported native speakers of English, and 30 for failing to pass the attentional checks (ACs) included in the experiment. We included two ACs: one at the very beginning of the experiment (before starting the training phases), and a second one before starting with the testing phase. According to our preregistration, participants who did not pass both attentional checks did not contribute to our sample.

[30] Both hypotheses above predict Condition to influence the pattern of responses for the held-out category, but they differ on the directionality of the predicted effect. The p-value of the model intercept indicates whether the log odds of selecting form 'A' over 'B' is significantly different from 0 for Condition 2 (the baseline). The coefficient itself indicates in which direction: given the coding scheme above, positive values indicate more extrapolation towards a 2+3 pattern, and negative towards a 1+2 pattern. While Harley and Ritter (2002) predict a significant negative coefficient, Ackema and Neeleman (2013, 2018) predict a non-significant coefficient. The p-values for the two fixed effects indicate whether responses in Conditions 1 and 3 are different from the baseline. Participants in both Conditions 1 and 2 are predicted to infer a 1+2 paradigm, therefore Harley and Ritter (2002) predict only Condition 3 to be different from the baseline. By contrast, Ackema and Neeleman (2013, 2018) predict both Condition 1 and 3 to be different from the baseline.

The model revealed that participants in Condition 2 (baseline) were significantly above chance in selecting 'form A' for the held-out category ($\beta = 4.15$; p$< .001^{***}$). In this condition, the trained meaning of form 'A' was 3rd person, and the held-out category was 2nd, therefore this result confirms that participants inferred a 2+3 homophony pattern. In Condition 3, the trained meaning of form 'A' was 2nd person, and the held-out category was 3rd. Therefore if participants in this condition are not significantly different from the baseline, this would suggest they inferred 2+3 homophony to the same degree as participants in Condition 2. This is confirmed by the model ($\beta = 0.614$; p$= .047$). By contrast, the model revealed a significantly lower proportion of 'form A' responses for held-out items in Condition 1 compared to the baseline ($\beta = -4.8$; p $< .001^{***}$). In this condition, the trained meaning of form 'A' was 2nd person, and the held-out category was 1st. Therefore we can conclude that participants did not infer 1+2 homophony to the same degree that participants in Conditions 2 and 3 inferred 2+3 homophony.

These results were further confirmed by two separate, intercept-only models for Conditions 1 and 3. For Condition 3, the proportion of form 'A' responses was significantly above chance ($\beta = 3.78$; p$< .001^{***}$), but for Condition 1, it was not ($\beta = -0.72$; p$= .29$). As in Condition 2, participants in Condition 3 inferred a 2+3 system.[31] However, there was no consistent pattern of inference in Condition 1.

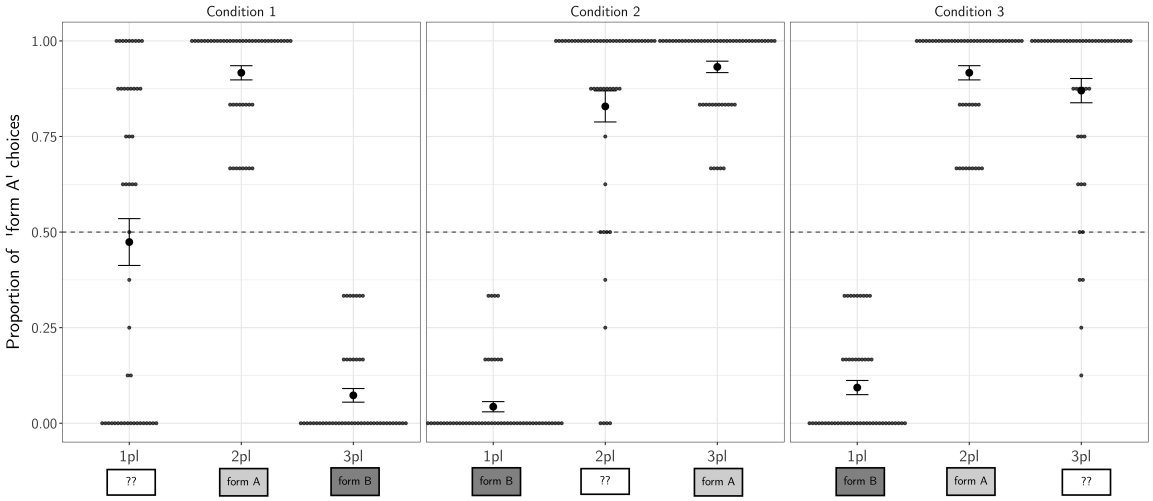


Figure 9: Proportion of form 'A' choices by condition during the testing phase. Choice of the same form across categories indicates homophony. Error bars represent standard error on by-participant means; dots represent individual participant means.

[31] This is further confirmed by a debrief questionnaire where participants were asked to provide information about the meanings of each of the pronouns. Most participants in Conditions 2 and 3 reported that the 2nd and 3rd person meanings were mapped into the same form.

4.3 Discussion

Results from Experiment 3 show that participants consistently inferred paradigms that feature 2+3 homophony whenever their training was compatible with this (Conditions 2 and 3). That is, learners preferred systems that collapse the 2nd and 3rd person categories to alternative homophony patterns. When 2+3 homophony was not available to participants (Condition 1), there was no stable pattern of responses: some participants used the same form for 1st and 2nd, some used the same form for 1st and 3rd, and others alternated randomly between the two. This result is in direct contrast to the prediction we derived from Harley and Ritter (2002). This theory is designed to derive 1+2 homophony, and *not* 2+3 homophony. Further it does not distinguish 2+3 from 1+3 homophony. Ackema and Neeleman (2013, 2018) fare slightly better, as their theory posits that 2nd and 3rd person form a natural class (DIST) that excludes the 1st person. This predicts that 2+3 homophony patterns should be preferred over 1+3 patterns, in accordance with our results (Forchheimer, 1953; Bobaljik, 2008, see also binary features accounts:[]). However, by also positing a feature shared between 1st and 2nd (PROX), this theory predicts no asymmetry between 2+3 and 1+2 homophony, contrary to our findings.

Thus, while the pattern of results in our experiment are quite clear, they do not straightforwardly match the predictions of either account. However, they do mirror what one might deem the most obvious typological asymmetry in table 8: paradigms featuring 2+3 homophony are the most common (see Table 9). Interestingly, an analogous pattern is found in the distribution of person systems across sign languages, where second and third person are consistently homophonous in both pointing signs and in other grammatical constructions (Meier, 1990; Neidle et al., 2000). This match between our results and typology suggests the possibility that some additional force is at play, which distinguishes 1st and 2nd person, even if they do form a natural class.[32]

To summarize, recall that at issue here was the special status of the ±participant distinction in driving homophony patterns. Theories based on the traditional set of binary features, like Harley and Ritter (2002), are designed to capture this particular natural class, thus predicting systematic homophony between speech act participants, i.e., 1st and 2nd person. Our results do not bear this out. Rather, behaviour appears to be consistent with

[32]It is worth briefly discussing two alternative interpretations of our experimental results. First, it could be that the little sister was treated as a speech-act participant rather than as a third person referent—for example because she is spatially close to the speaker and hearer (Peter Ackema, p.c.). To address this possibility, we re-ran one condition of this experiment with a different set of images where the little sister was seated in the background together with the parents. The results replicate the findings reported here, suggesting this is not an issue (see Appendix for full results; Figure 11).

Second, participants might think the speaker (i.e., the girl with the hat) is in some cases directly addressing the "others" (e.g., parents) instead of the intended hearer (the girl asking the questions). If an *addressee-shift* of this sort were at play, there would be no 'true' third person, explaining why participants derive what seems to be a 2+3 pattern (as suggested by Klaus Abels, p.c.). However, in the post-experiment debrief, most participants (across all three conditions) use a third person pronoun to describe the meaning of the form referring to others as 'them', and not 'you'. This strongly suggested they do not interpret the speaker as directly addressing the others.

a preference to distinguish the speaker from others. This distinction was discussed in the context of Zwicky's generalization and Experiment 2 above. We therefore return to this issue in the General Discussion below.

5 General discussion

The experiments reported above aimed to bring behavioral evidence from learning to bear on how person systems are represented, and whether they are subject to universal constraints on possible partitions. We see our results as making three main contributions: (i) we provide confirmatory evidence that the person space is represented in terms of primitive features, (ii) we point to the need for restrictions on patterns of inclusive homophony based on weak biases rather than hard-and-fast constraints, (iii) we provide new evidence for an asymmetry between 2+3 and other (non-inclusive) homophony patterns. We first summarize each of these in turn, and then discuss a number of broader issues raised by our results as a whole.

In Experiment 1, we sought to provide evidence that the person space is represented as an interacting set of universal features. While this idea forms the basis of all theories of person in the theoretical linguistics literature, to-date the link to learnability has been left implicit. However, these theories make testable predictions: if learners represent the person space as the interaction of a (primitive) set of person features, then categories which form a natural class should be readily mapped to the same phonological form (as formalized in Pertsova, 2011, among many others). The learner must simply determine which feature(s) are underspecified. A set of cells which cannot be characterized as constituting a natural class arguably requires learners to first learn the different categories (i.e., as feature bundles) and then independently learn to map the relevant set to a phonological form. Thus a feature-based theory of the person space predicts that paradigms with this kind of random homophony should be less readily learned than homophony among meanings with a shared feature. By contrast, if learners perceive the person space as a simple conjunction of four person categories (first, inclusive, second and third), then there is no basis for predicting that some homophony patterns should be more learnable than others.

The predictions of feature-based approaches to person were borne out in Experiment 1: homophony patterns among forms sharing a feature were indeed easier to learn, and more likely to be inferred when extrapolating a known form to a new meaning. This was the case even though the specific features involved were not active in participants' native language, suggesting that this is how these systems are represented. These findings draw an obvious parallel with recent work in phonology, showing that rules that characterized based on features (or natural classes) are easier to learn than rules which group together a set of featurally distinct segments (e.g., Saffran and Thiessen, 2003; Cristià and Seidl, 2008; Moreton and Pater, 2012, among many others). This suggests that feature-based representations are relevant for learning across domains, although the set of relevant features is domain-specific.

In Experiment 2, we turned to Zwicky's classic typological observation about person partitions without an inclusive distinction: they assimilate inclusive with the first person (rather than the second or third). Different approaches to person treat this apparently strong asymmetry differently. Harbour (2016) and Ackema and Neeleman (2018) are purpose-built to derive first+inclusive homophony only (though Ackema and Neeleman, 2018, may also be able to generate second+inclusive homophony).[33] However, approaches like Harley and Ritter (2002) can derive both first+inclusive and second+inclusive (though the latter is more marked).

Again, cashing out the predictions of these theories in terms of learnability, we tested whether homophony of the inclusive with first, second, or third were treated distinctly by learners. We found that third-inclusive homophony was particularly problematic, while first+inclusive had only a weak advantage over second+inclusive. In accordance with Harley and Ritter (2002), this suggests that there is no hard constraint against second/inclusive, rather inclusive can form a natural class with both the first and second, though not (readily) with the third.

Finally, in Experiment 3, we provided new evidence for an learning-based asymmetry between partitions with homophony between second and third person (2+3) on the one hand, and both first+second (1+2) and first-third (1+3) homophony patterns on other other. While the former were readily inferred by our participants, the latter were avoided. This result roughly matches an asymmetry found in the typological distribution for pronominal systems. However, it is problematic under the assumption that all homophony patterns targeting meanings that share a feature should be equally possible. Indeed, theories which posit that first and second person form a salient natural class, of speech act participants, Harley and Ritter, 2002), would predict that 1+2 homophony should if anything have a learnability advantage. Alternative approaches such as Ackema and Neeleman (2018) posit common features between both first and second categories, and between second and third, therefore the most straightforward prediction from this is a specific learnability disadvantage for 1+3 homophony. Participants in Experiment 3 instead consistently inferred 2+3 whenever this was consistent with the input they were trained one, while 1+2 and 1+3 were equally dispreferred.

To summarize, our results reveal that learners represent the person space in units which are smaller than person categories (features) and that instantiate natural classes. Natural class-based similarity therefore clearly plays an important role in determining how humans partition this person space. Indeed, a bias towards patterns based on natural class similarity pushes learners to preferentially collapse categories that share features (when they are required to do so). Our results support the claim that inclusive and second person should be included in the set of natural classes in this domain. However, there is also reason to believe that a bias for keeping the speaker distinct may also be at play. Indeed, our findings are compatible with the idea that natural class-based similarity, and speaker- (or

[33] Ackema and Neeleman's system, however, do allow some second-inclusive system to arise (violating Zwicky's generalization) whenever (a) PROX and PROX-PROX have different phonological realisations, and (b) there is an impoverishment of DIST in the plural when it's a dependent of PROX.

ego-)based distinctiveness may be two independent forces influencing the learnability of person systems. In Experiment 2, the partition characterized by first-inclusive homophony was learned most readily and is consistent with both of these pressures: it involves homophony among meanings which share a feature, +speaker, *and* keeps person categories implicating the speaker distinct from other categories. The second+inclusive homophony pattern is the next best, involving homophony among shared meanings, but not maintaining speaker categories as distinct from others. In Experiment 3, these two pressures led learners to infer 2+3 whenever they can—again among the options presented to learners in this case, 2+3 homophony is the only one to both involve categories that are highly similar and keep the speaker category distinct. The fit is not perfect: in principle, as noted above, we might have expected the first+inclusive advantage to be stronger, and 1+2 homophony to be preferred over 1+3. One possibility is that our failure to find these differences might be the result of particular features of our experimental design. For example, a difference in learnability between 1+2 and 1+3 might be revealed in an ease of learning experiment (cf. Experiment 2), where resorting the the preferred 2+3 pattern is not possible.

In our discussion of Experiment 2, we suggested that a speaker distinctiveness bias may be the result of the cognitive importance of the ego. As mentioned above, research on early pronoun acquisition has argued for an egocentric bias, whereby children perceive the world as a function of their presence on it and adopt categorisation systems that carry this distinction (Charney, 1980; Loveland, 1984; Moyer et al., 2015). This is also supported by the fact that perspective-taking appears to be a capacity that takes time to develop; infants and young children do not always succeed on so-called theory of mind tasks which require them to recognize that their internal knowledge states are not the same as other people's (e.g., see Ruffman, 2014). If the pressure for keeping the speaker distinct in pronoun systems comes from a general egocentric bias, then we might expect (i) the bias to be stronger in children, and (ii) the bias to be stronger when learners are active speakers of the language. The experiments reported here are of course on adults, and our participants were never themselves the speaker. Further research could, however, test both predictions.

An alternative is to build a speaker-based asymmetry between natural classes directly into a theory of person—at the same level of representation as the domain-specific primitive features. In a nutshell, the idea would be that the natural class 'speaker' is somehow represented as special within the person space. The pressure for speaker distinctiveness would then be a special type of natural-class-based similarity that comes not directly from the set of primitive features but from the particular status of the speaker feature. This sketch is very speculative, and a worked out implementation is beyond the scope of this paper. However, this idea is in the spirit of Harbour (2016), whose theory encodes an inherent asymmetry between speaker and addressee, built into the ontology. Crucially, we would argue that this asymmetry should be treated as a bias, which shapes but does not strictly delimit the space of possible partitions.

6 Conclusion

Person systems have been extensively explored from a theoretical standpoint: a number of approaches have been proposed, each of which constrains the set of possible person partitions that humans can represent, with the aim of explaining the prevalence of certain partitions of the person space cross-linguistically. The set of experiments reported here inform these theoretical approaches by generating direct behavioral evidence for the impact of hypothesized representations and constraints on the learnability of different person partitions. Indeed, our results constitute the first experimental evidence for learnability differences in this domain.

Specifically, we have provided evidence that there is a universal basis for a set of primitives organizing the person space that learners are sensitive to regardless of their native language. Looking more closely into the nature of these primitive features, we have shown that a theory of the person space needs to account for the semantic similarity between inclusive and second persons, on the one hand, and between second and third person categories, on the other. Each of these pairs of categories were treated as natural classes by learners, suggesting that they have features in common (e.g., an addressee feature possibly shared between second and inclusive persons). Besides a preference for feature-based patterns, there was evidence across our experiments that participants have an additional bias towards partitions of the person space where the speaker is distinct from other categories. Thus even though both speakers and addressees are participants in the conversation, there is an inherent asymmetry in how learners treat them. We sketched two possible accounts of this, the first in terms of a general egocentric bias, and the second in terms of a special type of natural class-based similarity.

More generally, our results suggest that the experimental paradigms developed here provide a fruitful method to test theoretically-motivated questions about how languages carve up the person space. While these kinds of methods have gained traction in investigating learning biases in a number of linguistic domains, here we have highlighted the sparsity of typological data as underscoring the need for new sources of evidence in building theories of person.

References

Ackema, P. and Neeleman, A. (2013). Person features and syncretism. *Natural Language & Linguistic Theory*, 31(4):901–950.

Ackema, P. and Neeleman, A. (2018). *Features of Person: From the Inventory of Persons to Their Morphological Realization*, volume 78. MIT Press.

Ackema, P. and Neeleman, A. (2019). Processing Differences Between Person and Number: A Theoretical Interpretation. *Frontiers in Psychology*, 10.

Baerman, M. and Brown, D. (2013). Syncretism in Verbal Person/Number Marking. In Dryer, M. S. and Haspelmath, M., editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Baerman, M., Brown, D., and Corbett, G. G. (2005). *The Syntax–Morphology Interface: A Study of Syncretism*. Cambridge University Press, Cambridge.

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

Barwise, J. and Cooper, R. (1981). Generalized quantifiers and natural language. In *Philosophy, Language, and Artificial Intelligence*, pages 241–301. Springer.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014). Lme4: Linear mixed-effects models using Eigen and S4. *R package version*, 1(7):1–23.

Bejar, S. (2003). *PHI-SYNTAX: A THEORY OF AGREEMENT*. PhD thesis, University of Toronto, Toronto.

Benveniste, E. (1971). *Problems in General Linguistics*, volume 8. Univ of Miami Pr.

Bickel, B. (2008). A refined sampling procedure for genealogical control. *Language Typology and Universals*, 61(3):221–233.

Boas, F. (1911). Introduction to the handbook of North American Indians. *Smithsonian Institution Bulletin*, 40:1–83.

Bobaljik, J. D. (2008). Missing persons: A case study in morphological universals. *The Linguistic Review*, 25(1-2).

Brown, C. A. (1997). *Acquisition of Segmental Structure: Consequences for Speech Perception and Second Language Acquisition*. PhD Thesis, McGill University Montreal, Canada.

Carstensen, A., Xu, J., Smith, C., and Regier, T. (2014). Language evolution in the lab tends toward informative communication. In *CogSci*.

Charney, R. (1980). Speech roles and the development of personal pronouns. *Journal of child language*, 7(3):509–528.

Chemla, E., Buccola, B., and Dautriche, I. (2019). Connecting Content and Logical Words. *Journal of Semantics*, 36(3):531–547.

Chomsky, N. (1965). *Aspects of Theory of Syntax*. MIT press.

Clements, G. N. (1985). The Geometry of Phonological Features. *Phonology Yearbook*, 2:225–252.

Corbett, G. G., Brown, D., and Baerman, M. (2002). The Surrey Syncretisms Database. http://epubs.surrey.ac.uk/816901/.

Cowper, E. and Hall, D. C. (2009). Argumenthood, pronouns, and nominal feature geometry. *Determiners: Universals and variation*, 147:97–120.

Cristià, A. and Seidl, A. (2008). Is infants' learning of sound patterns constrained by phonological features? *Language Learning and Development*, 4(3):203–227.

Culbertson, J. (2012). Typological Universals as Reflections of Biased Learning: Evidence from Artificial Language Learning: Typological Universals as Reflections of Biased Learning. *Language and Linguistics Compass*, 6(5):310–329.

Culbertson, J. (toappear). Artificial language learning. In *Oxford Handbook on Experimental Syntax*, page 26. to appear edition.

Culbertson, J. and Adger, D. (2014). Language learners privilege structured meaning over surface frequency. *Proceedings of the National Academy of Sciences*, 111(16):5842–5847.

Culbertson, J., Gagliardi, A., and Smith, K. (2017). Competition between phonological and semantic cues in noun class learning. *Journal of Memory and Language*, 92:343–358.

Cysouw, M. (2003). *The Paradigmatic Structure of Person Marking*. OUP Oxford.

Cysouw, M. (2010). On the Probability Distribution of Typological Frequencies. In Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Mattern, F., Mitchell, J. C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M. Y., Weikum, G., Ebert, C., Jäger, G., and Michaelis, J., editors, *The Mathematics of Language*, volume 6149, pages 29–35. Springer Berlin Heidelberg, Berlin, Heidelberg.

Cysouw, M. (2013). Inclusive/Exclusive Distinction in Independent Pronouns. In Dryer, M. S. and Haspelmath, M., editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Cysouw, M. A. (2005). Quantitative methods in typology. In *Quantitative Linguistik: Ein Internationales Handbuch= Quantitative Linguistics*, pages 554–578. de Gruyter.

Daniel, M. (2005). *Understanding Inclusives. Clusivity, Ed. by Elena Filimonova, 3-48*. Amsterdam: John Benjamins.

Dunn, M., Greenhill, S. J., Levinson, S. C., and Gray, R. D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 473(7345):79–82.

Fedzechkina, M., Jaeger, T. F., and Newport, E. L. (2012). Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, 109(44):17897–17902.

Finley, S. and Badecker, W. (2009). Artificial language learning and feature-based generalization. *Journal of Memory and Language*, 61(3):423–437.

Forchheimer, P. (1953). *The Category of Person in Language*. Walter de Gruyter GmbH & Co KG.

Futrell, R., Hickey, T., Lee, A., Lim, E., Luchkina, E., and Gibson, E. (2015). Cross-linguistic gestures reflect typological universals: A subject-initial, verb-final bias in speakers of diverse languages. *Cognition*, 136:215–221.

Gärdenfors, P. (2004). *Conceptual Spaces: The Geometry of Thought*. MIT press.

Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., Gibson, M., Piantadosi, S. T., and Conway, B. R. (2017). Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences*, 114(40):10785–10790.

Goldin-Meadow, S., So, W. C., Ozyurek, A., and Mylander, C. (2008). The natural order of events: How speakers of different languages represent events nonverbally. *Proceedings of the National Academy of Sciences*, 105(27):9163–9168.

Greenberg, J. H. (1988). The first person inclusive dual as an ambiguous category. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, 12(1):1–18.

Halle, M. and Marantz, A. (1994). Halle-Marantz1994.pdf. *MIT working papers in linguistics*, 21:275–288.

Hanson, R. (2000). 022_Winter_00_Hanson.pdf.

Hanson, R., Harley, H., and Ritter, E. (2000). Underspecification and universal defaults for person and number features. In *Actes de l'ACL/CLA Conference Proceedings. Ottawa: Cahiers Linguistiques d'Ottawa*, pages 111–122.

Harbour, D. (2008). On homophony and methodology in morphology. *Morphology*, 18(1):75–92.

Harbour, D. (2011). Descriptive and explanatory markedness. *Morphology*, 21(2):223–245.

Harbour, D. (2014). Paucity, abundance, and the theory of number. *Language*, 90(1):185–229.

Harbour, D. (2016). *Impossible Persons*. Linguistic Inquiry Monographs. The MIT Press, Cambridge, MA.

Harley, H. (2008). When is a Syncretism more than a Syncretism? Impoverishment, Metasyncretism, and Underspecification. In Harbour, D., Adger, D., and Béjar, S., editors, *Phi Theory: Phi-Features across Modules and Interfaces*, number 16 in Oxford Linguistics. Oxford University Press, Oxford ; New York. OCLC: ocn162126786.

Harley, H. and Ritter, E. (2002). Person and Number in Pronouns: A Feature-Geometric Analysis. *Language*, 78(3):482–526.

Hupp, J. M., Sloutsky, V. M., and Culicover, P. W. (2009). Evidence for a domain-general mechanism underlying the suffixation preference in language. *Language and Cognitive Processes*, 24(6):876–909.

Ingram, D. (1978). *Typology and Universals of Personal Pronouns. Universals of Human Language, Vol. III. Word Structure, Ed. by Joseph H. Greenberg, 213–248*. Stanford, CA: Stanford University Press.

Katzir, R. and Singh, R. (2013). Constraints on the lexicalization of logical operators. *Linguistics and Philosophy*, 36(1):1–29.

Kay, P. and Regier, T. (2007). Color naming universals: The case of Berinmo. *Cognition*, 102(2):289–298.

Kemp, C. and Regier, T. (2012). Kinship Categories Across Languages Reflect General Communicative Principles. *Science*, 336(6084):1049–1054.

Kemp, C., Xu, Y., and Regier, T. (2018). Semantic Typology and Efficient Communication. *Annual Review of Linguistics*, 4(1):109–128.

Ladd, D. R., Roberts, S. G., and Dediu, D. (2015). Correlational Studies in Typological and Historical Linguistics. *Annual Review of Linguistics*, 1(1):221–241.

Loveland, K. A. (1984). Learning about points of view: Spatial perspective and the acquisition of 'I/you'. *Journal of child language*, 11(3):535–556.

Martin, A. and Peperkamp, S. (in submission). Sensitivity to phonetic naturalness in phonological rules: An investigation of learning and sleep consolidation.

Martin, A., Ratitamkul, T., Abels, K., Adger, D., and Culbertson, J. (2019). Cross-linguistic evidence for cognitive universals in the noun phrase.

McGinnis, M. (2005). On Markedness Asymmetries in Person and Number. *Language*, 81(3):699–718.

Meier, R. P. (1990). Person deixis in American sign language. *Theoretical issues in sign language research*, 1:175–190.

Moreton, E. (2008). Analytic bias and phonological typology. *Phonology*, 25(1):83–127.

Moreton, E. and Pater, J. (2012). Structure and Substance in Artificial-phonology Learning, Part I: Structure. *Language and Linguistics Compass*, 6(11):686–701.

Moyer, M., Harrigan, K., Hacquard, V., and Lidz, J. (2015). 2-year-olds' comprehension of personal pronouns. In *Online Proceedings of the 29th Annual Boston University Conference on Language Development*, page 11, Boston.

Neidle, C. J., Kegl, J., Bahan, B., MacLaughlin, D., and Lee, R. G. (2000). *The Syntax of American Sign Language: Functional Categories and Hierarchical Structure*. MIT press.

Nevins, A., Rodrigues, C., and Tang, K. (2015). The rise and fall of the L-shaped morphome: Diachronic and experimental studies. *Probus*, 27(1).

Noyer, R. R. (1992). *FEATURES, POSITIONS AND AFFIXES INAUTONOMOUS MORPHOLOGICAL STRUCTURE*. PhD thesis, MIT.

Pagel, M., Atkinson, Q. D., and Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature*, 449(7163):717–720.

Pertsova, K. (2011). Grounding Systematic Syncretism in Learning. *Linguistic Inquiry*, 42(2):225–266.

Piantadosi, S. T. and Gibson, E. (2014). Quantitative Standards for Absolute Linguistic Universals. *Cognitive Science*, 38(4):736–756.

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

Piantadosi, S. T., Tenenbaum, J. B., and Goodman, N. D. (2016). The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological Review*, 123(4):392–424.

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

Rodrigues, A. D. (1990). You and I = Neither You nor I: The personal system of Tupinamba. page 13.

Rouse, S. V. (2015). A reliability analysis of Mechanical Turk data. *Computers in Human Behavior*, 43:304–307.

Ruffman, T. (2014). To belief or not belief: Children's theory of mind. *Developmental review*, 34(3):265–293.

Saffran, J. R. and Thiessen, E. D. (2003). Pattern induction by infant language learners. *Developmental Psychology*, 39(3):484–494.

Saldana, C., Oseki, Y., and Culbertson, J. (2019). Do cross-linguistic patterns of morpheme order reflect a cognitive bias? page 7.

Sauerland, U. and Bobaljik, J. D. (2013). Syncretism Distribution Modeling: Accidental Homophony as a Random Event. page 27.

Schwartz, L. J. and Dunnigan, T. (1986). Pronouns and pronominal categories in Southwestern Ojibwe. *Pronominal systems*, pages 285–322.

Silverstein, M. (1976). Feature hierarchies and Ergativity. *Grammatical categories in Australian languages*, pages 112–171.

Sokolovskaja, N. K. (1980). Nekotorye semantičeskie universalii v sisteme ličnyx mestoimenij. *Teorija i tipologija mestoimenij*, pages 84–103.

Sonnaert, J. (2018). *The Atoms of Person in Pronominal Paradigms*. LOT, Netherlands Graduate School.

Steinert-Threlkeld, S. and Szymanik, J. (2019). Ease of Learning Explains Semantic Universals. *Cognition*, page 21.

Tabullo, á., Arismendi, M., Wainselboim, A., Primero, G., Vernis, S., Segura, E., Zanutto, S., and Yorio, A. (2012). On the Learnability of Frequent and Infrequent Word Orders: An Artificial Language Learning Study. *Quarterly Journal of Experimental Psychology*, 65(9):1848–1863.

Wechsler, S. (2010). WHAT 'YOU' AND 'I' MEAN TO EACH OTHER: PERSON INDEXICALS, SELF-ASCRIPTION, AND THEORY OF MIND. *Language*, 86(2):332–365.

Werker, J. F. and Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant behavior and development*, 7(1):49–63.

White, J. (2017). Accounting for the learnability of saltation in phonological theory: A maximum entropy model with a P-map bias. *Language*, 93(1):1–36.

Wilson, C. (2006). Learning Phonology With Substantive Bias: An Experimental and Computational Study of Velar Palatalization. *Cognitive Science*, 30(5):945–982.

Zaslavsky, N., Kemp, C., Tishby, N., and Regier, T. (2018). Color naming reflects both perceptual structure and communicative need. *arXiv:1805.06165 [cs]*.

Zwicky, A. M. (1977). Hierarchies of person. In *Papers from the... Regional Meeting. Chicago Ling. Soc. Chicago, Ill.*, volume 13, pages 714–733.

7 Appendix

	Training phase I		*Training* phase II		*Testing* phase
Block	Exposure	What if...	Exposure	What if...	Testing
Person categories	3 SG	3 SG	All (8)	All (8)	All (8)
Speech roles switch	✓	–	✓	✓	—
Trials	12	6	16	32	24

(a) Experiment 1a

	Training phase I		*Training* phase II		*Testing* phase
Block	Exposure	What if...	Exposure	What if...	Testing
Person categories	3 SG	3 SG	All (8)	All (8)	All (8)
Speech roles switch	✓	–	✓	✓	—
Trials	12	6	16	32	24

(b) Experiment 1b

Table 10: Procedure in Experiments 1A and 1B.

	Control phase		*Critical* phase			
Block	Exposure	Testing	Exposure	Testing	Exposure	Testing
Person categories	3 SG	3 SG	4 PL	4 PL	4 PL	4 PL
Speech roles switch	✓	✓	✓	✓	–	–
Trials	12	12	16	16	8	32

Table 11: Experimental procedure in Experiment 2. Only participants who responded accurately to more than 2/3 of the trials in the Control phase could move forward to the Critical phase.

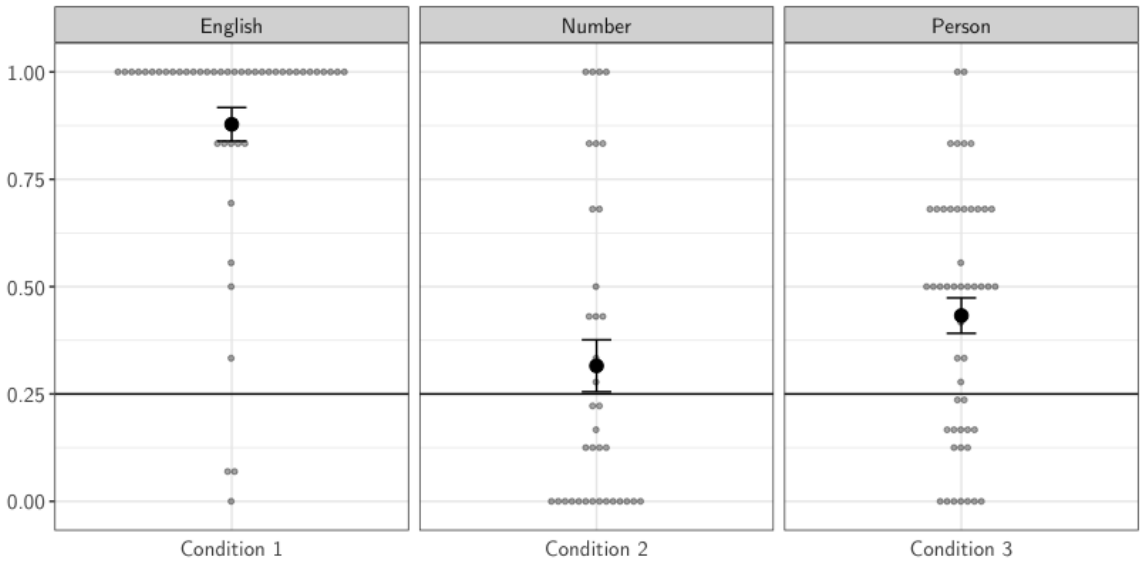


Figure 10: Probability of the feature-based pattern per subject and per condition in Experiment 1B.

	Training phase I		*Training* phase II		*Testing* phase
Block	Exposure	What if...	Exposure	What if...	Testing
Person categories	2 SG	2 SG	2 SG + 2 TR	2 SG + 2 TR	2 SG + 2 TR + 1 HO
Speech roles switch	✓	–	–	✓	✓
Repetitions	SG×4	SG×2	SG×2; TR×2	SG×2; TR×8	SG×2; TR×8; HO×8
Trials	8	4	8	20	24

Table 12: Experimental procedure for Experiment 3. After the training phase I, participants were asked to type in the meanings in English for the two words they had been exposed to. They are were given feedback on their responses, ensuring for the rest of the experiment that they knew the words were pronouns.

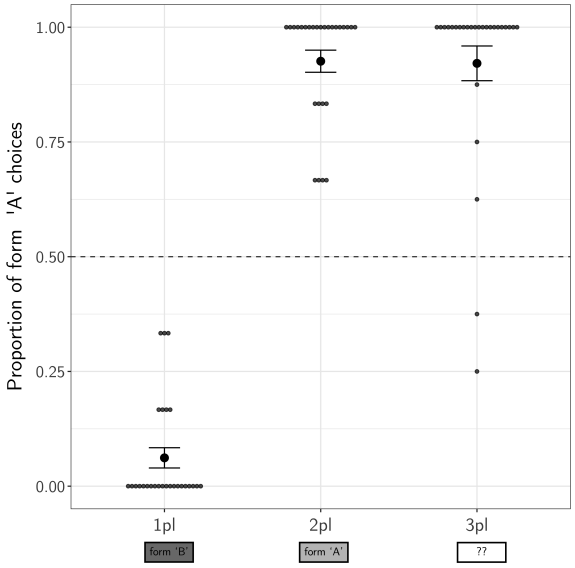


Figure 11: Replication of Condition 3 in Experiment 3 for group with little sister in the background (N=27, after exclusions).