frontiers

Social salience discriminates learnability of contextual cues in an artificial language

Péter Rácz [1,2,*], **Jennifer B. Hay** [2,3] **and Janet B. Pierrehumbert** [2,4,5]

[1] *Department of Archeology and Anthropology, University of Bristol, United Kingdom*
[2] *New Zealand Institute of Language Brain and Behaviour, University of Canterbury, Christchurch, New Zealand*
[3] *Department of Linguistics, University of Canterbury, Christchurch, New Zealand*
[4] *Oxford e-Research Centre, University of Oxford, United Kingdom*
[5] *Department of Linguistics, Northwestern University*

Correspondence*:
Péter Rácz
Department of Archeology and Anthropology, University of Bristol, United Kingdom, peter.racz@bristol.ac.uk

ABSTRACT

We investigate the learning of contextual meaning by adults in an artificial language. Contextual meaning here refers to the non-denotative contextual information that speakers attach to a linguistic construction. Through a series of short games, played online, we test how well adults can learn different contextual meanings for a word-formation pattern in an artificial language. We look at whether learning contextual meanings depends on the social salience of the context, whether our players interpret these contexts generally, and whether the learned meaning is generalised to new words. Our results show that adults are capable of learning contextual meaning if the context is socially salient, coherent, and interpretable. Once a contextual meaning is recognised, it is readily generalised to related forms and contexts.

Keywords: salience, language variation and change, experimental linguistics, morphology, indexicality, sociolinguistics, artificial language learning

1 INTRODUCTION

Studies of sociolinguistic variation show that people are able to associate linguistic patterns with a wide array of non-linguistic contexts (see e.g. Hay and Drager 2007; Drager 2010). What remains unclear is how these associations are learned, and whether learners discriminate these contexts in some structured manner. This learning problem is central in the sense that it sheds light on both the way contextual linguistic variation is structured and the way adults acquire it during their lifetime.

Of particular interest in this paper is the degree to which learners may attend differently to different types of non-linguistic context. Does the social salience of the non-linguistic context affect success in associative learning?

Context is interpreted in a number of ways in the relevant literature. For sociolinguists, the non-linguistic context is very broad. It includes the addressee and the discourse situation, as well as the speaker's attitudes and ideologies, which are conjoined to give social meaning to a given utterance (Eckert, 2008). For psycholinguists and psychologists, the context can also encompass higher-level situational attributes. In a given experiment, however, it can have a more specific interpretation, such as the visual field (Chun and Jiang, 1998) or the speaker (Kraljic et al., 2008b).

Salience is also interpreted in a number of ways, even within linguistics (Rácz, 2013). The core meaning is bottom-up and perceptual (a salient entity differs from its environment). We contrast this meaning of salience with *social salience*, a top-down, phenomenological concept, which encompasses the observer's background knowledge on the relevance of various aspects of the interaction (hence the term 'social'). We will use the concept of salience to differentiate non-linguistic contexts that are equally complex in structural terms but are used to different degrees in anchoring linguistic variation. We use it as a general, neutral term for this distinction between contexts.

In this paper we introduce an experimental paradigm that facilitates investigation into the contextual learning of morphological patterns. Using this paradigm, we then conduct a series of six experiments that together demonstrate a very significant effect of social salience upon contextual morphological learning.

1.1 Denotative and social meaning

Linguistic constructions have denotative meaning and social meaning. Broadly speaking, the former is the concept that the construction denotes, while the latter is the social-cultural context of its use.

Denotative meaning does depend on the context – the denotative meanings of even common concrete nouns can vary with the topic of discussion and the use of metaphor; *bug* means something different in discussions of gardening and of computer programming. However, the social and topical dimensions of word choice are only moderately correlated (Altmann et al., 2011).

Social meaning – for example, information conveyed about who is speaking, who is being addressed, and the nature of their relationship – is more indirect and more variable than denotative meaning (Foulkes and Docherty, 2006; Preston, 1996; Labov, 2001; Silverstein, 2009). Both Labov and Silverstein note that awareness to social connotations can vary from having explicit stereotypes to showing no social interpretation for social *indicator* variables that correlate with specific contexts in a way that is not acknowledged by the speakers.

In addition to dialect (Wells, 1982) and social group (Eckert, 2000), robust factors that influence social variation in language include age, gender, and sexual orientation (Tagliamonte and Roeder, 2009; Labov, 2001; Pierrehumbert et al., 2004). An important aspect of the social context is the addressee, or the speaker's relationship with the addressee (within its context). Linguistic accommodation to the addressee is a well-researched phenomenon (Soliz and Giles, 2014; Coupland et al., 1991). Certain languages, like Djirbal, develop lexical sets that are used with addressees belonging in specific kinship groups (Dixon, 1980).

Social meaning also varies in even more nuanced ways, as speakers dynamically use social-contextual information to take stances and negotiate in social interactions – different linguistic choices reflect the individual's linguistic experience and construction of social identity (Milroy, 1980; Eckert, 2000).

Listeners can, in turn, use such patterns to infer speaker characteristics or to adapt to different speakers when processing speech (for review of relevant literature see Hay and Drager (2007); Foulkes and Hay (2015)). Some words are statistically associated with older speakers and others with younger speakers; sensitivity to these associations can be displayed through varied ease of psycholinguistic processing without any explicit awareness of age-based patterns on the individuals' part (Walker and Hay, 2011). Listeners are able to associate different speaker personae with different combinations of morphological and phonological variables, based on fine-grained patterns in the ambient language (Campbell-Kibler, 2011). Individuals can also shift their categories of speech sounds based on cues of the broad cultural context, even if these cues are only peripherally present (Hay and Drager, 2010).

These examples show that social meaning can be attached to linguistic constructions that are not specific words or phrases – it can be generalised across linguistic patterns. It can apply to contexts of various levels of abstractness and expand to new contexts. It encompasses a wide variety of linguistic detail, from phonetics to word choice (German et al., 2013; Säily and Suomela, 2009). It relies on some contextual differences more heavily than on others, and speakers use it in a complex and subtle way to express intent and create a public persona (Eckert, 2012).

While numerous sociolinguistic and anthropological studies have revealed the importance of social meanings in language, less is known about how they are learned. Our understanding of social cognition in language leaves many unanswered questions about what details are noticed and remembered (Silverstein, 2009), as well as about what factors support generalisation to new forms or new situations (Pierrehumbert, 2006). Foulkes (2010:6) laments the lack of understanding of learning and storage of social meaning, stating that: ‘it now seems uncontroversial to conclude that social information is retained in memory alongside linguistic knowledge. Questions remain, however, over what sorts of social information are learned and stored, where and how they are stored in relation to linguistic information, and how social information affects linguistic processing.’

1.2 The role of the context in learning linguistic categories

If context is important, how do we learn to use it?

The contextual learning literature that is most relevant to this paper focusses particularly on the role of the broader extra-linguistic context – above and beyond the referent – in learning a linguistic category. Three important findings emerge from it.

First, consistency across contexts aids recall: a category is remembered more accurately in the context in which it was originally learned. In word-learning tasks, words are retrieved more accurately if recall occurs in the same location as training. Godden and Baddeley (1975), for example, show that words learned underwater are more accurately retrieved underwater. This relates to more general work on memory retrieval, where it has been shown many times that consistency of contextual information between encoding and retrieval leads to increased recall. Smith and Vela (2001) review literature showing that ‘people tend to be aware of their surroundings even when memorising something. As such, environmental features are typically encoded along with the to-be-remembered material’. Recurrent information will also invoke the context it was acquired in. Models of category learning and memory retrieval (Ratcliff, 1978; Grossberg, 1987) operate using notions of context-specificity. Certain memories activate specific contexts in which they were learned.

Broad experimental work in psychology discusses the role of contextual cues in category learning (Chun and Jiang, 1998; Goujon et al., 2015). This work shows that visual decision tasks speed up if the trial (with a given visual context) is shown repeatedly, despite the fact that participants are unable to identify the contexts afterwards, suggesting that the effect of the context on learning can be implicit. Qian et al. (2014) show that in a ‘whack-a-mole’ type game, players are faster at predicting the location of the mole if the location is probabilistically cued by moving background images that the player is not overtly oriented to. Gómez (2002) note that a consistent structure is learned better across multiple contexts. Observing the same pattern across multiple speakers improves learning as well (Rost and McMurray, 2009). Individuals use contextual memory to aid recall and prediction. Lleras and Von Mühlenen (2004) revisit Chun's paradigm, and their results indicate that the success of contextual cues depends on whether participants are focussing on the task in a narrow sense (presumably discarding the context) or are trying to take a holistic approach.

Second, categories learned in one context can generalise to another, similar context. Van der Zande et al. (2014) show that phonetic categories that shift due to exposure to a speaker retain this shift even when listening to another speaker. Maye et al. (2008) extend this generalisation to a new accent: their participants are able to use a context-specific vowel adaptation mechanism to process phonetic variation coming from a speaker and then, in turn, re-use the adapted vowel categories when encountering a similar speaker. Kraljic and Samuel (2006) show that a phonetic category distinction is generalised to a new linguistic context and also to a new speaker. That is, even if participants are trained on the distinction in one speaker voice, they carry over the distinction to another voice.

Third, listeners do not treat all available information the same way. Kraljic et al. (2008b) find that, while we first learn all phonetic detail as characteristic of a given speaker, we are later able to re-assess this knowledge and discard contextual variation that is based on an arbitrary idiosyncrasy of the speaker (such as talking with a pen in the mouth). Kraljic et al. (2008a) show that phonetic variation is processed differently if it is due to a consistent idiosyncrasy of a speaker (like a speech impediment) than if it represents a dialectal contextual allophone. Leung and Williams (2012) show that a distinction based on the animacy of referents is learned much more easily and generalised more readily than a distinction based on size differences between referents. Similar results have been found even for purely phonological contexts. Becker et al. (2011) find that Turkish speakers apply some statistical regularities in the Turkish lexicon, but not others, in a forced-choice wugs task; they conclude that speakers distinguish accidental from well-grounded statistical generalizations.

We are able to associate linguistic categories with non-linguistic contexts, even if these contexts are fairly arbitrary. We can extend this knowledge to similar but different contexts as well. And yet, we do not rely on all differences the same way – we distinguish information that is *relevant* in the context from information that is *accidental* to it.

1.3 Weighing the social salience of contexts

The amount of detail observed in sociolinguistic variation (Hay and Drager, 2007), coupled with memory models (Nosofsky, 1988), suggests that language users are able to construct social meaning based on a vast number of contexts. Some assume, however, that human memory is too restricted for this. Therefore, at least some of the information may be discarded if it cannot be used to make generalisations and if it taxes resources overtly. (For the debate, cf. Baayen et al. 2011; Gluck and Myers 2001; Denton et al. 2008).

How do we choose between useful and irrelevant contextual information? While the statistical co-occurrence of contexts and patterns is important, this is not the complete story. Selective attention guides which details are more important in processing information (Itti et al., 1998). Variation can be interpreted differently depending on its source (Kraljic et al., 2008a).

Relevance in turn derives from complicated assumptions about how the world works: that some speaker differences are consistent and others are haphazard, and that some contexts are more informative of language variation than others. When these assumptions include assumptions about what sorts of people are members of the same group, they rest on social constructs or categories. What information is grouped together and what is discarded both play a role in structuring social-contextual language variation.

Many experimental studies have explored the associations between familiar social groups and accentual features. Relatively few studies have investigated learning involving novel social groups, or learning of socio-linguistic cues other than ones at the phonetic level. This may be because it is a daunting task to set up scenarios in which the relationship between the social context and the linguistic pattern is transparent and well-controlled. However, there are several noteworthy studies. Work by Docherty et al. (2013) and Langstrof (2014) finds that people can associate familiar dialectal variables with arbitrary 'tribes' in a laboratory setting. Roberts (2008) shows that people are able to come up with morphological markers in a nonce language in order to demarcate in-group and out-group membership in a laboratory setting. Beckner et al. (2016) find that participants shift their linguistic patterns to accommodate to a group of human peers but not to a group of humanoid robots. The Beckner study shows extension of the accommodation pattern to new words that are similar in form to previously encountered ones. These studies do not look at extension or generalisation to different speakers.

The term *relevance* often implies conscious decision making on which contexts to consider and which to discard. Work on phonetic learning, however, suggests that we discriminate contexts largely implicitly. We will use the term *social salience* to compare the 'usefulness' of a context for linguistic learning and – as a consequence – people's ability to rely on it.

1.4 Individual variation in contextual learning

As in any learning task, people's success rates will vary in contextual learning. Work on language variation and change provides an important body of evidence on how people learn linguistic patterns that are associated with non-linguistic contexts. Labov (2001) shows that a new sociolinguistic variant does not diffuse uniformly through the population. Typically, there are community leaders of language change, who are chiefly responsible for spreading an innovative variant in the community. Later work distinguishes components of this process, all of which are relevant here. First, computational models have led to the conclusion that an innovative variant succeeds only if it carries a positive social weight, something which of course depends on learning a social association for the variant in the first place. (cf. e.g. Baxter et al. 2009; Fagyal et al. 2010). However, this positive weight does not need to be present in the minds of everyone in the community, but only in the minds of a critical group of early adopters – people who take up and use the new variant before other people do (Pierrehumbert et al., 2014). Experimental evidence for the existence of linguistic early adopters is found in Schumacher et al. (2014), an artificial language learning experiment in which some participants adopted an unexpected number-marking system far more than others.

These results indicate that individual variation in contextual learning is far from being a footnote example of differences in task completion – it is significant for patterns of language variation and change.

The source of such individual differences, then, becomes crucial, but it is less clear. Some individuals are likely better at recognising or remembering contextual patterns than others. For verbal and non-verbal statistical learning tasks, Siegelman and Frost (2015) show that individual performances in a verbal or a non-verbal task are not strongly correlated with different measures of intelligence and cognitive capacity. In fact, Siegelman and Frost find little correlation across performance in different statistical learning tasks. The analysis of Lleras and Von Mühlenen (2004) of individual participant behaviour in their learning experiments (adopted from the work of Chun and Jiang 1998) indicates that participant success in a task depends on the strategy adopted by the participant. This high-level, complex decision is unlikely to be derived from any single cognitive or linguistic measure. However, some systematic effects have been identified. Vocabulary size is a good predictor of how easily new word forms are learned in children (Henderson et al. 2015 and therein). Henderson et al. (2015) do not find an effect in adults, but in a related pseudo-word rating study with more statistical power, Needle et al. (2015) do find that high-vocabulary adults have a better ability to decompose nonce compounds such as *angstroof*. Brooks et al. (2016) shows that learning and generalizing an L2 morphological pattern can be partially predicted by measures of non-verbal intelligence and statistical learning ability.

Ramscar et al. (2014) argue that the life-long accumulation of experience affects performance in psycholinguistic tasks. Older people have more prior experience, and so work with a denser cue space in verbal tasks. In an explicit learning task with feedback, Metcalfe et al. (2016) finds that older participants perform better, especially on unfamiliar items. Event-related potentials for the older participants indicate better ability to focus attention on feedback.

In short, participant accuracy is highly variable in learning tasks and this variation derives from a complex set of cognitive differences. However, two studies (Ramscar et al., 2014; Metcalfe et al., 2016) point to age as a interesting factor. Prior experience may affect performance on psycholinguistic tasks, either via richer mental representations gained through experience, or though better proficiency in allocating attention. Recruiting participants on Amazon Mechanical Turk provided us with a participant pool of diverse ages that makes it possible to assess this factor.

Diverse sources of evidence indicate that language use relies on contextual cues, and that speakers evaluate these cues both implicitly (based on salience and statistical co-occurrence) and more explicitly (based on social salience). How this behaviour is learned is less clear, but the learning process and the individual differences manifested in it are both very relevant to the study of language variation and change.

2 AIMS

In this paper, we report on a series of experiments that build on previous results in context-specific category learning. In the experiments, participants have to learn linguistic patterns that depend on the context. The context can be linguistic – the choice of a suffix depends on the shape of the stem. It can be non-linguistic – the choice of a suffix depends on the conversation partner. Conversation partners can differ across various dimensions. Both the patterns and the contextual differences are more abstract than those explored in studies of phonetic adaptation such as Kraljic and Samuel (2006).

The linguistic patterns we look at are morpho-phonological. They are suffixation patterns in a simple artificial language that mark the diminutive or the plural. These are both transparent, iconic relations that also show considerable variation in English and other languages. The specific linguistic contextual pattern we use (the suffix vowel should match the stem vowel) is not found in English, so participants must learn it. Their success in doing so provides the baseline condition for the experiment, serving to validate the paradigm and shed light on the strengths of the effects found for the various social factors.

The social dimension we focus on is socially robustly interpretable, the gender of the conversation partner. We contrast it with a dimension that has similar visual prominence but lacks its social salience, the spatial orientation of the conversation partner. We chose to explore the gender distinction because it is a very robust sociolinguistic marker. Children as young as 6 months, for example, preferentially match sex-cued voices and faces (Walker-Andrews et al., 1991). Sex and gender have a complex effect on the use of social meaning in general (Cheshire, 2002; Milroy and Milroy, 1993). Our experiments build on each other to provide a solid foundation for the salience of this distinction, by showing that it holds up across differing amounts of exposure, types of extensions, or types of linguistic patterns.

We ask the following questions:

1. Given a morpho-phonological pattern, how quickly and well are participants able to learn its association with a linguistic context? Are they able to generalise the pattern to new words? How does generalisation to new words compare to words seen in training?
2. How well can participants associate a morpho-phonological pattern with a social context: conversation partner gender? Are they able to generalise the pattern both to new words and to new instances of the appropriate context?
3. Are all types of social association equally learnable and generalisable? Or is an incidental social property (spatial orientation) processed differently from a more static conversation partner characteristic (gender)?
4. How do individuals vary in learning contextual associations for linguistic patterns?
5. Are older participants more successful, or less successful, at learning such social contextual associations?

As we will see, for the morpho-phonological patterns we look at, learning is possible for both linguistic and non-linguistic contexts. For non-linguistic contexts, participants are more successful in learning an association with a robust, salient context, conversation partner gender. This context is interpreted broadly – gender is recognised as the defining dimension. The morphological pattern is recognised and extended to previously unseen words after training. Older participants are better learners in our data.

We find these results with more types of conversation partners (such as children and adults) and with two distinct morphological patterns, the diminutive and the plural. These results indicate that the salience of conversation partner gender *vis-à-vis* spatial orientation is a broad and general phenomenon.

3 OVERVIEW OF EXPERIMENTS

Our experiments use a training-test paradigm based on a simplified version of adaptive tracking. The adaptive tracking paradigm described in Leek (2001). It was previously adapted to linguistic research by Schumacher et al. (2014). We discuss the paradigm in detail in the Methods Section of Experiment I.

Experiment I trains participants on a morphological pattern, presented visually. They see picture pairings, with a large and a small version of the same entity. The large version is named. They have to choose the name of the small version, which is a suffixed form of the name of the large version. There are two suffixes, and the correct one includes a vowel that matches the stem vowel. We find that participants learn to consider this context easily and also extend the pattern to new items.

In Experiment II the same morphological pattern is presented, again with different conversation partners. This time the pattern depends on the conversation partner. There are two conversation partners - a male and a female - and both are presented in two ways visually, in side view and in front view. One group of participants is trained with the morphological pattern depending on conversation partner identity (answers for conversation partner A or B pattern together) The other group is trained with the pattern depending on conversation partner spatial orientation (answers pattern together according to whether the conversation partner is presented in front view or side view).

We find that an association of the pattern with the identity of the conversation partner is easier to learn than an association of the pattern with the spatial orientation of the conversational partner. This result resembles the findings by Kraljic et al. (2008b) on learning of incidental versus characteristic patterns of phonetic variation. The morphological pattern is interpreted broadly – in the test session, it is extended to items not seen in the training session.

Experiment III expands the scope of contextual learning to examine whether participants generalize on the basis of conversation partner gender. Gender is one of the most widely discussed predictors of sociolinguistic variation. Morpho-phonological and lexical variation that depends on gender is not restricted to languages like Dyirbal. Languages like French and German use different adjective conjugations depending on the referent, including first and second person referents in the discourse, while stochastic differences for gendered language use have been found in English as well (Hay and Walker, 2013). We find that gender is a better cue than spatial orientation. Naming is readily extended to new conversation partners that fit this context (i.e. as another female or male conversation partner).

Experiment IV focuses on the way participants rely on the denotative and the contextual aspect of the naming pattern. The general layout is similar to experiments I-III. However, the test phase is different. Instead of a right and a wrong answer, they are forced to choose between one answer that is correct in its denotative aspect but wrong in its contextual aspect and another one that is set up the other way around. We find that if the contextual cue is conversation partner *gender*, participants have split preferences between the denotation and contextual cue. With a spatial cue, they overwhelmingly prefer the denotative aspect.

In experiments V and VI we extend the paradigm to investigate a new morphological pattern – the plural – and investigate the effect of a radically increased training set size. We find that learning the plural is similar to learning the diminutive. Increased training improves participant accuracy in test.

4 EXPERIMENT I

Experiment I establishes our experimental paradigm and investigates the role of the linguistic context in learning within this paradigm.

In Experiment I, participants learn a morphological pattern that is sensitive to the linguistic context. It is a vowel harmony or partial reduplication pattern (common in the world's languages, though not present in English): the vowel of the suffix has to match the vowel of the suffixed stem. The version of adaptive tracking presented here was used successfully by Schumacher et al. (2014), who also recruited participants on Amazon Mechanical Turk. Our design, however, differs in its overall theme, as well as in the amount of training participants receive.

We explain our design in detail in Section 4.2 and address changes to it in subsequent Methods sections.

4.1 Participants

The experiment was hosted on Amazon Mechanical Turk (AMT). 47 people participated in the experiment. 22 are women, 25 men. All are native speakers of American English. We base this claim on the fact that all participants had IP addresses from the United States and self-identified as native speakers (those who did not were excluded from the results). The mean age is 31 years, with a standard deviation of 8.5. Participants were paid three dollars upon completion of the task.

For each of our experiments, we used Amazon Mechanical Turk worker IDs to exclude participants who had taken part in any of the other experiments. US worker IDs are independently verified by Amazon, making it very difficult for the same person to operate multiple accounts.

As in all following experiments, we used training speed to remove outliers (cf. below). For each across-subject condition, we removed the 2.5% of participants who took the most trials to finish training – i.e. the slowest ones. We filtered participants within the across-subject conditions, since we expect conditions to vary in length. In Experiment I, which has one across-subject condition, we removed 2 participants and report data from 45 participants. We return to the training phase and discuss our exclusion criteria in detail in the Results section of Experiment I.

For each experiment, we do not report the precise ratio of AMT workers who picked up the task versus workers who finished it, since an online task can be interrupted for various reasons, including connection issues, disruptions, etc. On the whole, about 5% of the workers who started these experiments did not complete them. This, in our experience, is not an excessively high number for an online experiment.

By using Amazon Mechanical Turk, a burgeoning forum for psycholinguistic research (Munro et al., 2010), we were able to recruit a large number of participants in a short span of time. Amazon Mechanical Turk is especially fit for our experiment, which has a 'game-ified', button-input design. The game format allows for immersion of the participants, and increases the likelihood that they pay attention to the task – otherwise they cannot finish it. *Gamification* has been increasing in popularity in data collection in recent years (Von Ahn, 2006) and it has been used successfully in linguistic experiments as well (Fedzechkina et al., 2012; Schumacher et al., 2014). Relying on Amazon Mechanical Turk allowed us to run substantial numbers of subjects, so as to be able to see important differences across conditions in how likely our various predictions are borne out. Crump et al. (2013) show that more complex laboratory tasks on category learning can be replicated using subjects on AMT. However, AMT subjects are, overall, less successful learners than laboratory participants, possibly because they are less focused and attentive when participating from their homes, without the presence of an experimenter.

Experiment I tests for a main effect of a single across-subject condition while Experiment II has two across-subject conditions and tests for interactions as well. This is why the latter has twice as many participants as the former. The same logic was applied to subsequent experiments. This is an economical use of participants, but does have the restriction that power to estimate participant-level effects (here, gender and age) will vary across experiments. We return to this issue in Section 8, in which we estimate these effects on a merged dataset.

This and the following experiments reported in the paper have been overviewed and approved by the Institutional Review Board of Northwestern University and the Human Ethics Committee of the University of Canterbury. During the time of data collection, the experimenters were not affiliated with any other institutions.

4.2 Methods

In Experiment I, participants play a computer game in which they have to help a bird flying roof to roof to return to its nest. The game consists of a training phase, followed by a test phase. The targets are presented in the following way. For a given *target*, the participant sees a *conversation partner* who shows a *query picture* to the main character, the bird, along with the *prompt*, the *name* of the depicted object. The bird responds with a *response picture* and *two possible names*. The participant has to choose one of them. The response picture is always the *diminutive* version of the conversation partner's picture (depicted as a small or juvenile version of the conversation partner's picture). This implies that one of the two possible names is the correct name of the diminutive of the query picture. A *target* is the combination of an item (a query-response picture pair) with a conversation partner. Figure 3 shows the general layout with examples of the phases and the mechanics. Stimuli are *visual* only.

During training, targets are presented in a random order to each participant, and the participant has to give a correct answer for every target in order to move to the next target. If they give an incorrect response, they have to return to the previous target. The test phase does not use adaptive tracking. Here, targets are presented, again, in random order. The targets in the test phase include both the training items and previously unseen items. No feedback is given. Training takes place during the in-game day, test during the in-game night. Participants are also told when they enter the test phase. The way training relies on simplified adaptive tracking guarantees that each participant has responded to each stimulus correctly at least once before moving to the test phase. Unlike training protocols with a fixed number of trials, it provides an opportunity for participants who find the task difficult to improve by training for longer.

The name of the query picture is a nonce word with a CVC structure. These are drawn from a set of 12 syllables (cf. Table 1). Half of the name syllables have the <e> vowel, the other half the <a> vowel. The two possible names of the response picture are the name of the query picture plus one of two suffixes – *pek* and *pak*. These suffix syllables were selected such that one had <e> and one had <a>. The correct response always matches the vowel of the prompt. This echoes vowel harmony or partial reduplication systems commonly found in natural languages. Participants encounter six items in training and these six items plus six new items in test. These are items that do not occur in training.

We designed the stimuli with the following principles in mind: (i) the syllables should be distinctive; (ii) they should consist of a small set of frequent letters; (iii) they should be easy to pronounce for our participants, who are American English speakers; (iv) the consonant clusters in the two-syllable words should cue English word boundaries in a uniform manner. These are somewhat competing requirements but our aim was to provide a relatively optimal set that balances all of these considerations.

Out of the set of syllables, the two suffixes are randomly assigned for each participant. So are names of the individual objects, using the remaining set of twelve syllables. Six occur in training and then also in the test. Six occur only in test.

In all our experiments, participants come across four conversation partner images during the game.

The images we use for our conversation partners in experiments I-VI can be seen in Figure 1. We will refer to them in the paper using the labels *woman*, *man*, *girl*, and *boy*. Each figure has two perspectives, *front* and *side*, giving us 8 conversation partner images in total.

The particular images we used were designed to be matching in many respects, while still appearing visibly different according to gender and/or view. It is difficult to assess the degree to which we were successful with this aim, as human raters completing an explicit similarity rating task would be unable to avoid bringing their social knowledge to bear. However, in order to attempt an objective test that there were not strong visual differences between the different dimensions, we computed the Levenshtein distance between uniformly binned histograms of the greyscale versions of the images using Matlab (Mathworks, 2016). Histogram comparison is a common method in image processing (Pele and Werman, 2010).

The Levenshtein distance between the images is roughly similar. For the adult images, there is a slightly larger distance between the woman and the man than between the front and side views for either the woman or the man. But for the child images, the order of the distances is interwoven between the grouping factors, and the largest difference is between the boy front and side view, and the smallest difference is between the girl and the boy front views. (Figure 2 is a tile plot that shows image distances. Darker hue means smaller distance.)

This visual similarity metric patterns differently for the adult and child figures, but - as we will report in the following sections - none of our experiments reveal any difference in whether participants were trained on the adult or the child images (c.f. experiments III; V). It therefore seems very unlikely that patterns relating to image similarity are driving the behaviors that we observe.

The four conversation partner images in Experiment I are the *woman* and *man* figures in Figure 1, viewed from the *front* and the *side*. All items occurred with all conversation partners, giving a minimum of 24 trials in training and 48 trials in test. Who the conversation partner is has no bearing on the correct name selection in training, since the latter only depends on vowel quality. Linguistic context is relevant, non-linguistic context is irrelevant. The conversation partners will become relevant in Experiment II. The 'comic book' setup of the experiment allowed us to freely combine text with conversation partner images in all experiments.

In our experiments, the visual field of the experiment (the window in which it takes place on the user's computer screen) is the non-linguistic context. The words that occur in this visual field (written in the latin alphabet) constitute the linguistic context. We use the visual display to manipulate a classic sociolinguistic factor, the *addressee*.

4.3 Participant instructions

The experiments are designed to create a setting for linguistic or socio-linguistic learning that is controlled, yet still somewhat naturalistic. The task itself is made explicit, but the potential cues for the correct answers are not. Participants need to work out which cues to attend to from the potential cues available. These include the orthographic shapes of the word forms, the item pictures, and the conversation partner pictures (since the protagonist and the background are held constant). This makes the task harder. But it also makes it analogous to problems we encounter in language use, an issue that has received considerable attention in the literature of contextual language learning (cf. Yu and Smith 2007).

Participants receive written instructions at the beginning of the game. They are told that the bird is the protagonist ('our hero'), and that they need to help our hero return to its nest by flying from roof to roof. The hero will meet people who stand on the roofs and ask questions. The hero needs to answer the questions correctly in order to proceed. The questions are explained: the person names an object and shows our hero a smaller version of the object as well. Our hero has to guess what they would call the small object. It is explained that a second phase follows this first phase. In this phase, participants need to guess the names given to small objects, just like they did in the first part. They are asked to try to remember what the right answers were in the first part, and guess the right answer based on that.

4.4 Hypotheses

We hypothesised that participants would learn the association of the morpho-phonological pattern with the linguistic context and generalise it to new items. Based on related studies of phonetic learning, such as the study of an indexical allophonic pattern reported in German et al. (2013), we also predicted higher accuracy for test items seen in training than for unseen test items. We also evaluated age and gender as potential predictors of performance.

4.5 Results

Overall, the results show that many participants succeeded in learning and generalising. This outcome is reflected in the time course of training and in accuracy in the test phase.

Since the length of training depends on the participant's success at the learning task, training length is a good indicator of task difficulty. It is also a good indicator of participant attention and ability.

We use trial counts to express training duration. While individual trials vary in duration (there is no time limit on trial length, that is, people can spend as much time as they want on their decision), they do so to a modest degree (in Experiment I, mean (m) = 16 sec, standard deviation (sd) = 12 sec).

We prefer trial count to duration in time because the latter can be affected by user computer problems, server lag, and participant behaviour (taking a break, answering the phone, etc.).

In similar experiments the accepted norm is removing participants who are 2 or 2.5 standard deviations outside the overall mean. This method has its problems, as shown by Leys et al. (2013), who recommend mean absolute deviation instead. For our data, neither the standard deviation threshold, nor the mean absolute deviation threshold are applicable. We decided to use a percentage threshold since trial length in our experiments is not normally distributed, making standard deviation a poor measure of the distribution of participant trial count. The distribution starts at 24 (the minimum possible number of trials) and has a long right tail. Our 2.5% threshold is arbitrary. A method of outlier removal that relies on the 2 standard deviations threshold would remove about 5% of the participants from the experiment. Our method removes the slowest 2.5%. A participant cannot finish too quickly, and so the distribution of training trial counts has no left tail. We remove outliers to safeguard against participants with very poor attention.

In all experiments, we filtered participants within the across-subject conditions because we expected these to vary in length. We used a quantile threshold to remove participants in the right tail of the training trial count distribution. For every condition, we establish the 0.975 quantile threshold of the distribution of training trial counts. We exclude participants over this threshold. The number of participants removed for each experimental condition ranges between 1 and 2. Outliers for the separate conditions add up to the sum of outliers for each experiment.

In Experiment I, 2 out of 47 participants are over the 97.5% threshold.

Participants finish training much faster than a player would by chance (m = 43, sd = 18). Individual variation for trial counts is large. Participants recruited through Amazon Mechanical Turk vary more in their behaviour than would the college students recruited for a typical lab experiment.

Experiment I has one across-subject condition. Training speed in this condition is only informative inasmuch as it is, on average, much shorter than what we expect if participants were guessing. This shows that some form of learning is taking place in training.[1]

Accuracy in test depends on whether the item was seen in training. Figure 4 is a bean plot of participant accuracy in the test phase, grouped by whether the item was seen in training.

The bean plot shows the distributions of participant responses along the y axis. Mean accuracy is higher for items seen in training (seen items, right) than for items participants only encountered in test (unseen items, left). This is indicated by the long black horizontal bars. The *distribution* of mean subject accuracy rates, however, is also revealing. For items not seen in training, we see a very clear bimodal distribution, with most subject means centred either around .5 or near 1 (up to 1, actually, since it is impossible to have a *higher* accuracy rate than 1). A person whose accuracy is around .5 in a task that involves binary choices is effectively guessing. A person whose accuracy is 1 has done a perfect job. For items seen in training, there also appears to be a bimodal distribution, but the total mass of the upper mode is greater and more participants perform at accuracies around .6 to .7.

[1] Due to the use of the simplified adaptive tracking paradigm, a player who guesses randomly on all training targets would need 518 trials, on the average, to finish training. The high number is due to the fact that, in the adaptive tracking paradigm, an incorrect answer returns the previous trial, and an incorrect answer to that throws the participant back even further.

We used the R statistical computing environment for our analyses (R Core Team, 2016). We created our plots using `ggplot` (Wickham, 2009).

We stepwise fit a binomial mixed-effects regression model on the test data, using response to individual items (*correct* or *not correct*) as an outcome variable and *presence in training* and participant *age* and *gender* as predictors, with a participant grouping factor (random intercept). (Bates et al., 2012; Gelman and Hill, 2006). We used a random intercept for participants to account for participant-specific differences in variation. Since object-name pairings are generated on the fly, these are different for each participant. As a consequence, we did not need to model item-level variation (e.g. with an item random intercept), making our models computationally more effective.

For each regression model in this paper, we started with a fully specified model including all interactions and removed non-significant predictors one by one, testing for model fit using analysis of variance and the Akaike information criterion (AIC). Where a combined model was too complex we fit interactions of participant-level predictors (age and gender) and experimental conditions (cue type, item presence in training, etc.) separately. We only report the best model, which means that we exclude predictors that were not significant.

8 out of 498 participants (across the 6 experiments) did not disclose their age in the pre-test survey. When we tested age as a predictor, we re-fit models excluding these missing data and performed analysis of variance checks on these models to inform model selection. Models excluding participants with missing data were consistent with models fit on full data. For models for which age was justified as a predictor, the reported models exclude the few participants for which we have no age data. This model selection process assures that (i) we use all the available information in our models and that (ii) participant-level and experiment-level factors, along with their interactions, are tested in each experiment.

The best model for the test phase of Experiment I can be seen in Table 2. The model includes participant age. For Experiment I, all participants reported their age, and so no participants needed to be excluded on this basis.

The model shows that participants are more likely to pick the correct suffix in test if they have seen the item in training. Age is a significant predictor – older participants are more likely to give correct answers. This effect is not strong compared to presence in training, but it is robust and remains even if we remove margin values.

4.6 Discussion

The results of Experiment I confirm that, within the current design, many people are able to accurately learn a morpho-phonological pattern they were trained on. They are able to ascertain the triggering linguistic context and choose the appropriate answer. They are also able to generalise this pattern to items not seen during training. This remains true despite the relatively low number of training items with which the cue was presented. However, performance is somewhat better for test items previously seen in training.

Test behaviour follows training behaviour closely. Participants who finish training earlier are more likely to have a high accuracy in test.

Figure 4 shows that average participant accuracy in test has a bimodal distribution. Those participants whose accuracy is below the overall mean (at 0.74) have means clustered around 0.5 (equivalent to chance), while participants above the overall mean have means clustered towards 1 (equivalent to a perfect score). Based on this difference we can divide participants into ‘good learners’ and ‘poor learners’. This grouping is supported by the training data. If we compare training length for the 'good learner' participants (those with mean above the overall mean in test) and the 'poor learners', we find that the former finish training faster (as supported by a Wilcoxon rank sum test, $W = 56$, $p < 0.001$). Note that there are 19 good learners and 26 poor learners, suggesting that the task is relatively hard (with a 'passing rate' of 42%). These results may be compared to those in Becker et al. (2011), an experimental study of nonce words in Turkish in which vocalic cues to an alternation were found to be less learnable than a consonantal cue. Our results also show that many participants have difficulties learning a vocalic cue to an alternation.

The mean trial count of the 'good learner' group in training is 32. That of the 'poor learner' group is 51. Recall that training has 24 unique trials. If a participant does not find out the key to success in training (the stem vowel), and keeps guessing, but remembers every single guess and identifies it correctly afterwards, they will need about 36 trials to finish training on the average (since they have a 50% chance of guessing right in the first place, and only need to repeat half of the trials). If they keep guessing, they need 518 trials on the average. The mean of the 'poor learner' group is clearly between these values, suggesting that some rote learning did take place for this group (no participant needed 518 trials to finish), but it was not entirely efficient.

The 'good learner' / 'poor learner' distinction is post-hoc. Although we expected individual variability in learning, we did not hypothesize beforehand that listeners would fall into two clusters, with rather few 'intermediate learners' falling between the 'poor' and 'good' learners. One possible interpretation is that the good learners are people who became consciously aware of the relevant cue. Conscious learning, also described in the research literature as 'explicit learning' is generally faster than unconscious, or implicit learning (Goujon et al., 2015). While the good learners recognise the contextual pattern and simply apply it to all new items, some poor learners seem to perform rote learning as they repeatedly see training items. They do learn the correct suffixed form for some specific items, as evidenced by the greater number of participants who perform above chance in seen items, but they are not successful in generalising to new items. It is also possible, of course, that this distribution does not relate to an explicit/implicit learning distinction at all, but rather reflects the distribution of individual learner characteristics in our data-set. Brooks et al. (2016), for example, have shown that morphological learning and generalization varies across individuals, in a way that correlates with measures of non-verbal intelligence and general statistical learning abilities. What follows is that if key individual learner characteristics are bimodal, then the learner outcomes would also be bimodal.

The results of Experiment I give us indications on how participants proceed through a learning task based on a cue associated with a linguistic context. The decisive point is whether a participant learned the pattern, and if this does not happen in training, participants will mostly guess in test. A sizeable group, but still a minority, learned the general cue association pattern. In Experiment II, we look at a similar task that uses a non-linguistic context.

5 EXPERIMENT II

In this experiment, the cue is no longer related to the name used by the conversation partner – rather, it is the conversation partner itself.

5.1 Participants

105 participants were recruited through Amazon Mechanical Turk. 51 are women, 54 men. Mean participant age is 34 years, with a standard deviation of 9.62. 54 participants were assigned to the *view*

condition, 51 to the socially relevant *gender* condition. Four participants were excluded for not following the instructions properly. Four participants were removed based on training speed. We report data from the remaining 97 participants. All participants are native speakers of American English. Each person was paid three dollars upon completion of the task.

5.2 Methods

Experiment II modifies Experiment I in one major way. The correct response no longer depends on the vowel of the prompt. Rather, it depends on the conversation partner, who was irrelevant in Experiment I. A non-linguistic contextual cue replaces the linguistic cue in learning a morpho-phonological pattern. The non-linguistic contextual cue is relatively basic. It is either who the conversation partner is or what physical orientation they have compared to the protagonist.

Experiment II has the same conversation partners as Experiment I, who, again, can each be seen in two different ways. This creates two groupings. One grouping, *gender*, is the identity of the conversation partner - who is either male or female. The other, *view*, is the spatial orientation of the conversation partner. The aim of this design is to teach naming patterns in conjunction with the *contextual cue* provided by the grouping. The images used can be seen in Figure 1. Learning the ’view’ cue requires participants to notice that changes in language use are correlated with changes in the direction the partner is facing. Learning the ’gender’ cue requires the participants to notice that changes in language use are correlated with changes in the speaker.

Who your conversation partner is has a huge effect on linguistic category learning. Listeners are able to keep track of information coming from two different speakers, adapt to new speakers, and recognise the difference between across-speaker and within-speaker variation and weigh them differently (Kraljic et al., 2008a; Horton and Gerrig, 2005, 2002). Perceived speaker gender is an especially robust cue (Johnson et al., 1999).

In contrast, conversation partner spatial orientation is much less salient as a social-indexical cue. People learn both deictic expressions (denoting spatial relations, such as ‘here’ and ‘that’) and words with implicit spatial relations (such as ‘wide’ or ‘tall’) easily, since these are frequent forms of every language. Variation between deictic expressions, however, does not typically carry social meaning.

Note that, if our participants in this task learn the association of the linguistic pattern with conversation partner, we have no way of knowing whether they are imputing a person-specific pattern, or a more general distinction based on person gender. As the most salient difference between the two partners is the gender difference, we here refer to the cue as a *gender* cue. Whether the learned cue is identity or gender can not be established from the design of Experiment II. However we will explicitly test the degree to which the learning to generalized to other speakers on the basis of gender in later experiments.

The game, then, has four conversation partner images. Each occurs once with each item in training. Again, targets are presented in a random order to each participant, and the participant has to give a correct answer for every target in order to move to the next target. In the test phase, targets are presented, again, in random order. No feedback is given. Training consists of six items, so it has 4x6=24 targets in total. The test contains these six items, and six items unseen in training, presented with each of the four conversation partners, so it has 4x12=48 targets in total.

The nonce language we used is similar to Experiment I, except that the vowel harmony pattern is absent. Instead, we used the five English vowel letters to make stimuli maximally distinct. The list of stimuli used in Experiment II can be seen in Table 3. The same principles guided stimuli selection as in Experiment I.

Since stem vowel is no longer relevant, we used the five English vowel letters to make the syllables more distinct. For each participant, two syllables are randomly selected as suffixes (marking conversation partner gender or spatial orientation, depending on the condition) while the rest are randomly assigned as item names.

There are four conversation partner images in the experiment, and two suffixes. Each suffix corresponds to two conversation partner images. The *across-subject* factor of Experiment I is the grouping of the conversation partner images. In the *gender* condition, the correct suffix (and, consequently, the correct response) is *cued* by the identity/gender of the conversation partner. In the *view* condition, the correct suffix (and so the correct response) is cued by the conversation partner's orientation (facing outwards or facing left).. The *within-subject* factor is whether a test item was seen in training.

5.3 Hypotheses

Experiment II looks at the association of a morpho-phonological pattern and a non-linguistic context. We had three hypotheses for Experiment II: (i) Participants would learn the diminutive pattern and extend it to new items in the *gender* condition (ii) learning and extension would be poorer in the *view* condition (iii) participants would be more likely to assign the correct pattern to items in the test phase if they have *seen* them in the training phase. We also evaluated participant age and gender as predictors of performance.

5.4 Results

We find that the pattern is indeed easier to learn with the *gender* condition. Unlike in Experiment I, item presence in training has no effect on response accuracy.

We use two measures of participant performance. In the training phase, we look at the number of trials it takes a participant to finish the experiment. This number provides information about the difficulty of learning in training and how much attention the participant pays to the task – this is why we use it as our main exclusion criterion. Participant responses in the test phase tell us how much they remember training and how easily they extend the pattern to new items and conversation partners.

Training takes longer (in terms of trial counts) in Experiment II (m = 66, sd = 25) than in Experiment I (m = 42, sd = 18) (a significant difference according to a Wilcoxon rank sum test, $W = 3450$, $p < 0.001$).

In Experiment II, participant training trial count is longer in the *view* condition (m = 74, sd = 27) than in the *gender* one (m = 59, sd = 22, $W = 1565$, $p < 0.01$). Training with the *gender* cue in Exp II is still significantly longer than training in Exp I. Figure 5 is a kernel density plot of training trial count for individual participants grouped by the two conditions. Mean trial count is shorter in both conditions than what we would expect for random behaviour. Trial count is the number of trials it takes a participant to finish training. The smoothing bandwidth and the y axis are held constant for all density plots in this paper to aid comparison.

Figure 6 is a bean plot of participant responses, contrasting the *gender* condition and the *view* condition. For the *view* cue, most participants have a mean around .5 – they are effectively guessing in test. In contrast, a sizeable proportion of participants has high accuracy for the *gender* cue. The bimodal structure of the *view* distribution strongly resembles the distribution of participant results in Experiment I for the unseen items.

We stepwise fit a binomial mixed-effects regression model on the test data, using response to individual items (*correct* or *not correct*) as an outcome variable and cue type (*gender* or *view*), *item presence in training*, and participant age and gender as predictors, with a participant random intercept. The summary for the best model can be seen in Table 4. Since age was a significant predictor, this model excludes 2 out of 97 participants who had no age data available. Model fitting in Experiment II is similar to Experiment I, we start with the most complex regression model and remove predictors one after the other until we reach the best fit. We test for all interactions of our terms.

The model shows that participants who are trained on the *gender* cue have much higher accuracy in test. Unlike in Experiment I, item presence in training is irrelevant – participant accuracy remains the same with previously seen and unseen items. Age is a significant predictor of test accuracy: older participants are more accurate.

5.5 Discussion

Cue type is a strong and independent predictor of test accuracy in Experiment II. Participants trained with the *gender* cue have a much higher test accuracy, echoing results in the contextual learning of phonetic categories. Item presence in training does not affect test accuracy.

The Somers' Dxy Rank Correlation between test response accuracy and training trial count is modest (0.37). This is probably because participants show two types of behaviour, much as in Experiment I. As we speculated above, some participants may have explicitly recognised the the context-pattern association, while others did not.

If we group participants with mean test accuracy above the overall mean as 'good learners' and those below the overall mean as 'poor learners', we find that good learners finish training in significantly fewer trials (W = 484, $p < 0.001$).

If we look at the distribution of good learners across cue type, we find that most good learners are to be found in the *gender* condition (cf. Table 5). This tabulation supports the results of the regression analysis: the context-pattern association is easier to recognise for the *gender* cue than for the *view* cue.

When we compare Experiment II with Experiment I, we see that learning the non-linguistic cue is harder than learning the linguistic cue. As we note above, training takes longer. This remains true if we compare the *gender* cue with the linguistic cue in Experiment I (learning the linguistic cue takes significantly fewer trials, W = 639, $p < 0.001$). An important difference between Experiment I and Experiment II, however, is that the linguistic cue is learned through exposure to a range of linguistic items, but the gender cue is learned through a contrast between just two people. A number of studies have shown that repetition and variability of context leads to improved learning (Gómez, 2002; Rost and McMurray, 2009). We cannot therefore directly compare the learnability of the linguistic and the social cue from these experiments alone.

Test accuracy for the *gender* cue in Experiment II is not worse overall than test accuracy in Experiment I. There is, however, an important difference in relation to item presence in training, which is significant in Experiment I but not in Experiment II. We merged the data from Experiments I and II and performed a binomial mixed-effects regression analysis, using the interaction of *item presence in training* and *cue type* (view, gender, linguistic) as predictors, with a participant random intercept. Effect sizes can be seen in Figure 7.

For the two contextual cues tested in Experiment II, item presence in training is not relevant. For the linguistic cue in Experiment I, participants are better at recalling names for items they have seen in training. This result could indicate that rote-based learning of items is relevant for the linguistic cue, but less so for the contextual cues (even for the *gender* cue, where a substantial amount of learning takes place). Note, however, that the role of the stem is different in the two experiments. In Experiment I, the participant must attend to the different stems in order to select the correct suffix. In contrast to Experiment I, the key to success in Experiment II is paying attention to the social context. The available responses always share the stem with the prompt word, which is irrelevant in relation to success on the task. This fact could explain the differing outcomes.

It remains clear that, in Experiment II, most learning happens with the socially salient, interpretable cue, *gender* – the identity of the conversation partner. The notion of social salience afforded by this experiment, however, is very narrow – it entails a distinction between two specific conversation partners (a woman and a man) as opposed to their position in space.

We have referred to the two cues as *view* versus *gender*, assuming it is very likely that participants rely on the visible gender difference between the conversation partners in making their decisions. In Experiment II it is impossible to know whether participants are performing a categorization based on speaker gender, or simply associating the cue with the particular speakers. In Experiment III, we therefore continue exploring non-linguistic contexts, by more explicitly testing whether the associations learned in this type of experiment are extended to other partners, on the basis of conversation partner gender.

6 EXPERIMENT III

In this experiment, we look at whether participants generalise from the learning process we have seen in Experiment II, by extending the contextual cue to new conversation partners on the basis of gender.

6.1 Participants

The experiment was hosted on Amazon Mechanical Turk. 101 people participated in the experiment. 57 are women, 44 men. 50 are in the *gender* condition, 51 in the *view* one. Mean age is 32 years, with a standard deviation of 10.83. Four were excluded from the analysis based on training length. We report data from the remaining 97 participants. All are native speakers of American English. Participants were paid three dollars upon completion of the task.

6.2 Methods

Experiment III was designed to replicate the results of Experiment II, and in addition it investigates whether participants are able to generalise the contextual cue to new conversation partners in the test phase. It was identical to Experiment II except for the fact that Experiment III has eight conversation partners instead of four. Four conversation partners are present in training and test (just like in Experiment II) and four conversation partners are only present in test. Both previously seen and novel items are presented with previously seen and novel conversation partners in test, making the test twice as long as in Experiment II.

Experiment III uses all the conversation partner images in Figure 1. Conversation partners are grouped according to a *gender* attribute, as well as a *perspective (view)* one, their spatial orientation. We used an adult/child distinction for conversation partner images that are present in training versus unique to the test. The reason for this is that we wanted to keep the two conversation partner categories distinct visually. Some higher level knowledge is needed to realise that an adult and a child share the same gender. In contrast, two adult images of the same gender could have been matched based on visual similarity only.

The experiment has two across-subject factors. Half the participants have to learn the relevant cue (*gender*), and half of them the accidental cue (*view*). Also, half the participants are trained with children, and the other half with adults, creating four different training groups.

6.3 Hypotheses

We evaluated four hypotheses for Experiment III: (i) Participants would learn the diminutive pattern and extend it to new items and new conversation partners in the *gender* condition, (ii) The diminutive pattern would be easier to learn if it is associated with the *gender* cue than with the *view* cue (iii) Participants would be better able to assign the correct pattern to items in the test phase if they have seen them in the training phase, (iv) participants would be better able to assign the correct pattern to conversation partners that they had seen in the training phase. We also evaluated age and gender as predictors.

6.4 Results

Participants finish training much faster than a player would at random. On average, it takes participants longer to finish training in the *view* condition than in the *gender* condition (a significant difference according to a Wilcoxon rank sum test, W = 1556, $p < 0.01$). Figure 8 is a density plot of training length for individual participants grouped by the two conditions. Training trial count in Experiment III is not significantly different from Experiment II.

Figure 9 shows a bean plot of participant test responses for the *gender* condition and for the *view* condition. Mean accuracy is much higher for the *gender* condition.

We stepwise fit a binomial mixed-effects regression model on the test data, using response to individual items (*correct* or *not correct*) as an outcome variable and the interaction of cue type (*gender* or *view*) and item *presence in training*, conversation partner *presence in training*, conversation partner *type* in training (*children* or *adults*), and participant age and gender as predictors, with a participant random intercept. Response accuracy is predicted by cue type. It does not depend on familiarity with items or conversation partners. Accuracy does not improve significantly with age, or differ by participant gender. The summary of the best model can be seen in Table 6.

6.5 Discussion

The results of Experiment III support the results of Experiment II, and show that the learning generalizes to other partners. Even when exposed to just one person in the training, participants extend this learning to others, on the basis of the person's *gender*.

While half the participants are trained with children and the other half with adults, this makes no difference in test accuracy.

This further supports our assumption that the perceptual difference between our conversation partner images is far less relevant than their socially salient grouping characteristics.

As in the previous two experiments, participant mean test accuracy ratings show a clear bimodal distribution. We can group participants as good learners or poor learners, according to whether their test mean is above or below the overall mean. If we tabulate good learners across cue type, we find that the *gender* cue is easier to learn. This can be seen in Table 7.

The results of Experiment III are very similar to Experiment II. The main difference is that, in Experiment III, we have evidence that participants clearly rely on a more abstract context to establish generalisations. If they recognise conversation partner gender as the contextual cue, they are able to interpret it generally. They are able to learn this cue with adults and extend it to children and vice versa. This is comparable to the recognition of phonetic categories in stereotypical male and female voices. The huge difference is, however, that this distinction is both much more abstract (relying on a distinction in diminutive use) and simpler (a single difference in suffixes as opposed to a complex envelope of distinction between stereotypical male and female voices). This grants additional power to our socially salient distinction, which is generalised to differences between stereotypically male and female characters. This distinction, trained with only one instance of each gender, is straightforwardly generalised to a new instance (from a woman to a girl, etc.).

Now that we have established that learning based on just one person is extended to another person of the same gender in the test, this substantiates the choice of *gender* (rather than identity) as the most appropriate label to use for the person-based cue.

Note that the item presence in training is not a significant predictor of test accuracy in either Experiment II or III. This suggests that participants completely disregard the prompt word form and focus on the suffix and the associated context (if they focus on anything at all). The design of these experiments, however, does not allow us to explicitly test whether participants pay attention to the suffix versus the stem and how this relates to training performance. Experiment IV addresses this question.

7 EXPERIMENT IV

In this experiment, we return to the learning process in Experiment II and look at the relative importance of our various cues by offering participants two test choices that are both ‘wrong’, in different ways.

7.1 Participants

The experiment was hosted on Amazon Mechanical Turk. 80 people participated in the experiment. 46 are women, 34 men. 40 are in the *gender* condition, 40 in the *view* one. Mean age is 32 years, with a standard deviation of 9.99. Two participants were excluded from the reported data based on training length. We report data from the remaining 78 participants. All participants are native speakers of American English. Participants were paid three dollars upon completion of the task.

7.2 Methods

Experiment IV uses the adult *woman* and *man* conversation partners in *front* and *side* view.

For Experiment IV, as for Experiment II, context determines the correct response during the training phase. Each target has two possible responses. One has the suffix associated with the present context, the other has the suffix associated with the absent context. So, if the context is *gender* the participant must choose the response with the suffix that matches the gender of the conversation partner on screen. The base of the two available responses is always the same, the name of the query, which is also visible on the screen.

The test phase of Experiment IV differs from Experiment II in two respects. First, during the test phase, participants are only exposed to previously seen items, no novel items are presented. And second, the query and the prompt name are no longer visible on the screen. One possible response for the target has the base which is the name of the query of the target (as seen in training) but a context-inappropriate suffix (this is a choice present in the previous experiments). The other possible response has the correct suffix, but it has a base that is not the name of the query of the target (as seen in training). Both answers are wrong (compared to training), but for different reasons. One has the correct prompt name, one the correct suffixation pattern, but neither has both. Table 8 gives an example.

In the training phase, the stimuli were generated from the same pool as in Experiment III. For each participant, two syllables are assigned as suffixes. Six syllables are assigned as item names. In test, the 'wrong conversation partner' answer was generated using the prompt name and the wrong suffix. The 'wrong prompt name' was generated using a different, randomly assigned prompt name and the correct suffix. This means that the wrong stems were different for the same item across test trials.

Picking the response with the correct base (the name of the query in training) but the wrong suffix (the one that belongs to the other cue) means that, during training, participants pay more attention to the entity they name than the context. Picking the response with the correct suffix (the one that belongs to the present cue) but the wrong base (not the name of the query) means that, during training, participants pay more attention to the context than the entity they name. The naming task in Experiment II and III derive the name from the query image and the conversation partner, and the naming task in Experiment IV allows directly comparing their degree of relevance.

7.3 Hypotheses

We had two hypotheses for Experiment IV: (i) as in the previous experiments, participants would finish training faster in the *gender* condition than in the *view* condition. (ii) Participants would be more likely to focus on the suffix in the *gender* than in the *view* condition; as seen in experiments II-III, the *gender* cue contributes more to learning success, and hence it is likely easier to recognise and learn.

7.4 Results

The overall training duration of Experiment IV is not significantly different from that of Experiment III or Experiment II. As in Experiment II, training in the *gender* condition is significantly shorter than in the *view* one ($W = 1023$, $p < 0.01$, using a Wilcoxon rank sum test).

In the test phase, overall, participants pick the answer containing the correct suffix significantly more often than the answer with the correct base (the 'original' name) (59% of the time).

During test, participants in the *view* condition pick the correct base overwhelmingly more (76% of the time) than in the *gender* condition (41% of the time). In the *gender* condition, the correct suffix is preferred more often (59% of the time).

Figure 10 shows the degree to which participants pick the base (1) or the suffix (0), that is, the preference for the *base* with the *view* cue (left) and the *gender* cue (right).

We stepwise fit a binomial mixed-effects regression model on the test results using 'picked correct base' (as opposed to 'picked correct suffix') as outcome variable and condition (*gender* or *view*) and participant age and gender as predictors, with a participant random intercept. The summary of the best model can be seen in Table 9. The only significant predictor is the condition, with the *view* cue leading to a stronger preference for the base than the *gender* cue.

7.5 Discussion

In the *view* condition, participants overwhelmingly focus on the base of the response, rather than the suffix. This suggests that, in the *gender* condition, the suffix is much easier to learn than in the *view* condition. This is also the outcome for Experiments II & III: accuracy for the *view* condition is not much higher than chance. Some participants learn the *view* cue, but many fewer than the *gender* cue.

The *gender* condition of Experiment IV is more interesting. The tight answer ratio (59% versus 41%) indicates that participants can rely on either – there are 'object' people and 'people' people. This difference does not vary with participant age or gender. We can infer more about being an 'object' or a 'people' person as a learning strategy if we look at training performance for these two groups. Figure 11 shows training trial counts for participants who overwhelmingly go for the base or the suffix in test.

People who go on to pick the suffix in test are much faster to finish training than people who go on to pick the stem. We can interpret training and test performance in Experiment IV as results of either of two learning strategies. 'Object' people focus on the stem, therefore take a while to finish the training, and overwhelmingly pick the stem in the test. 'People' people focus on the suffix, finish training much faster, and pick the suffix in the test. We should note that 'object' participants in the *gender* condition are as slow in training as participants in the *view* condition.

In the *gender* condition, as in the analogous conditions of experiments II-III, both the linguistic context and the non-linguistic context of the morphological pattern (the prompt name and the conversation partner's gender/identity) are readily available. A group of participants are able to pin down the relevant factor in variation, namely, the conversation partner, and pick their responses accordingly. Others remain 'distracted' by the prompt name. In the *view* condition, the non-linguistic context is barely if at all accessible – consequently, all participants focus on the linguistic context.

We use the word 'focus' to refer both the participant's attention (what part of the frame they pay attention to) and the participant's weighing of the cues (how much importance they attribute to any cue; that is, any part of the frame that changes from trial to trial). These cannot be separated in our analysis, but likely together lead to the observed dichotomised participant behaviour, dividing successful and poor learners in the experiment.

We now have strong reasons to believe that a robust and salient non-linguistic context is easier to learn than a less salient one. The generality of these findings, however, is somewhat compromised by the fact that we have thus far only looked at one morphological pattern, the diminutive, which is both highly variable in English and which has strong associations with gender in many languages (Jurafsky, 1993/2012). In order to make our findings more robust, we repeated Experiment III using the plural instead of the diminutive as the iconic relationship between prompts and targets. The main question was whether participant accuracy changes with visual stimuli cueing the plural replacing diminutive stimuli.

8 EXPERIMENT V

In this experiment, participants work with an artificial language that is based on a different iconic relationship, the plural instead of the diminutive.

8.1 Participants

The experiment was hosted on Amazon Mechanical Turk[2]. 89 people participated in the experiment. 50 are women, 39 men. 46 are in the *gender* condition, 43 in the *view* one. Mean age is 37 years, with a standard deviation of 15.47. All are native speakers of American English. Participants were paid three dollars upon completion of the task. We excluded 4 participants from the analysis based on training speed. We report data from the remaining 85 participants.

[2] We initially ran 40 participants in this experiment. A reviewer pointed out that the lower participant count is problematic given that we compare this experiment to Experiment III. We have then run additional participants for Experiment V. Regression analysis shows no difference in the performance of the first and second batch in Experiment V

8.2 Methods

Experiment V is identical to Experiment III except for the prompt and target images. Experiment III, like all previous experiments, used a normal sized item and a diminutive item as the pair of pictures. The instructions told the participant to identify the name of the small item based on the larger item. In Experiment V, by contrast, each query picture displays an item and each target picture displays three of the same item. The instructions tell the participant to identify ‘the plural, the word for multiple instances of the same item’. Otherwise instructions are unchanged. The goal of Experiment V is to determine whether the results of Experiment III generalise to a morphological process (pluralisation) that is highly general and productive in English and many other languages.

Experiment V uses all conversation partners in *front* and *side* view.

8.3 Hypotheses

Our hypothesis was that the patterns that we had previously observed would generalize beyond the particular case of the diminutive. We therefore evaluated the same hypotheses for Experiment V as for Experiment III. Based on the results of Experiment III, we expected that participants would learn the plural pattern and extend it to new items and new conversational partners. We expected learning to be more successful in the *gender* condition than in the *view* condition. We expected no advantage for seen items or partners, and we also looked for effects of participant age and gender.

8.4 Results

Here, we first look at Experiment V by itself and then together with Experiment III.

Cue type has no effect on training speed in Experiment V.

The mean rate of participant accuracy in test can be seen in Figure 12.

We fit a mixed-effects binomial regression model on the test data using item and conversation partner presence in training, type of cue, and participant age and gender as predictors, with a participant random intercept. The model summary can be seen in Table 10. The only predictor that shows any effect is cue type (β = 0.84, se = 0.46, p = 0.07). The effect size is above the level of statistical significance. This result is similar to what we see in Experiment III, even though the effect is weaker in test – and absent in training.

The only way to tell whether the experiments differ from each other significantly is to use statistical tests on the joint data from the two experiments.

Training in Experiment V does not differ significantly in length from training in Experiment III.

We merged the two datasets and stepwise fit a mixed-effects binomial regression model on the combined test data using item and conversation partner presence in training, type of cue, type of pattern (*diminutive* or *plural*), and participant age and gender as predictors, with a participant random intercept. The plural dataset patterns essentially the same as the diminutive dataset. *Cue type* is a significant predictor of test accuracy. Participant age and gender and item presence in training are not significant. The type of pattern (diminutive/plural) does not affect test accuracy significantly. The summary of the best model of the merged test data can be seen in Table 11.

8.5 Discussion

Experiment V shows that the learning difference between the socially salient cue and the irrelevant cue persists when these cues are tied to a different morphological pattern, the plural. This adds further robustness to this distinction.

When we look at the experiment in itself, the effect of the gender cue is weaker than in other experiments, e.g. Experiment III. It is also above the generally accepted threshold of statistical significance. However, the statistical analysis of the two datasets together indicates that this difference is not statistically significant. The joint analysis gives us no ground to reject the null hypothesis that the plural does not differ from the diminutive. In general, finding significance levels for differences in statistical significance is difficult and would require a study considerably exceeding our scope at present. If future work establishes that socio-indexical associations for plural patterns are indeed more difficult to learn than for diminutive patterns, the reasons for this difference would be of considerable interest. Potential factors could include adult differences in the adaptability of the derivational vs the inflectional morphology, and pre-existing associations between the diminutive and social attributes of age, gender, or status (as described in Jurafsky 1993/2012 as well as Kruisinga 1942 cited by Bauer 1997). For English, a further aspect is that the language has a number of competing diminutive suffixation patterns (such as *-ling, -ly, -ie*, etc), but only one broad, productive plural pattern.

In experiments II, III, and V, it is only with the salient cue that participants show a large degree of learning. However, only about half the participants exposed to the salient cue show high accuracy in test, while the other half of this group resorts to guessing, much like participants learning the non-salient cue. In Experiment IV we looked at learning strategies and proposed that, when both types of information are accessible, some participants will focus on the linguistic context (the prompt), and others at the non-linguistic context (the conversation partner). What remains unclear is whether participants make a by and large random choice at the beginning to focus on either context and then remain with it, or whether the effect of the non-linguistic context can be increased by expanding training. This is an especially relevant question given that item presence in training does not affect test accuracy, suggesting that the recognition of the relevant context is far more important than exposure to the specific training items.

One important question arising from our results across experiments I-V is the role of individual participant characteristics. We evaluated age and gender as individual predictors, with mixed results. These participant characteristics were not controlled in the participant recruitment procedure, and different experiments enrolled slightly different age and gender distributions.

In order to obtain more statistical power to look at these participant effects, we combined the test data for experiments II, III, and V, which have the same training size, the same test setup, and the same cue differences (*gender* and *view*). The pattern is either the diminutive or the plural. Those are also comparable. Each experiment has items in the test phase that were seen in training as well as new items. We fit a binomial mixed-effects regression model on the combined test data for participants with age available – 273 participants. The outcome variable is response accuracy, the predictors are participant age and gender, as well as cue type and item presence in training, with a participant random intercept. The best model has age (β = 0.03, se = 0.01, p $=$ 0.01) and cue type (β = 1.23, se = 0.25, p $<$ 0.001) as significant predictors. The model summary can be seen in Table 12. Older participants are more accurate overall, and responses to socially salient cues are much more likely to be correct. Cue type is a much stronger predictor than age. Participant gender is not a significant predictor. Similarly, inspection of the training data reveals a significant age effect, with older participants completing the training in significantly fewer trials.

The result shows that older participants are doing better with the socio-indexical learning. The effect, however, is not very strong.

9 EXPERIMENT VI

In this experiment, participants undergo an extended training phase. Extended training allows us to explore whether participants can modify their focus of attention based on feedback during training.

9.1 Participants

The experiment was hosted on Amazon Mechanical Turk. 100 people participated in the experiment. 55 are women, 45 men. 58 are in a *gender* condition. 42 participants are in the *view* condition. Mean age is 35 years, with a standard deviation of 10.98. Four participants were excluded based on training length. We report data for the remaining 96 participants. All participants are native speakers of American English. Participants were paid three dollars upon completion of the task.

9.2 Methods

Experiment VI is based on Experiments II and V. The morphological pattern is the plural, as in Experiment V. The extent of exposure is three times as great as in Experiment V: The plural pattern is trained with 18 (instead of 6) items and 4 conversation partners, and it is tested with these 18 items and 18 previously unseen items. There are no unseen conversational partners in the test phase, as in Experiment II. Since the focus is on the effect of familiarity with training items and since including new conversation partners as well would have prolonged the experiment to a large degree, we only included conversation partners seen in training in the test. We use the same conversation partners as in experiments I, II, and IV: the *woman* and the *man*. Our list of stimuli was expanded for Experiment VI (cf. Table 13). The main principle was to avoid adding syllables with consonant clusters which would upset the symmetry of the concatenated words. This was achieved by adding ‘ng’, the English consonant letter for the velar nasal, to the set of available syllable codas.

Experiment VI uses the adult *woman* and *man* conversation partners in *front* and *side* view.

9.3 Hypotheses

Our hypotheses were based on the the results of Experiments II and V. We expected that participants would learn the contextual association more easily in the *gender* condition than in the *view* condition. We also expected that they would generalize to new items. In Experiment VI, we were also seeking a more in depth understanding of individual success rates. In addition to evaluating the effects of participant age and gender, we asked whether the lengthened training phase improves the success rate, compared to the previous experiments, and whether it affects the distribution of the good learners versus poor learners.

9.4 Results

As in the previous experiment, we first analysed the data from Experiment VI and then went on to compare it with Experiment II, which has a similar setup but shorter training.

Similar to most previous experiments, training takes longer with the non-salient cue (here: *view*) than with the salient cue (here: *gender*) (W = 584, $p < 0.001$).

The mean rate of participant accuracy in test can be seen in Figure 13.

We fit a regression model on the test phase of Experiment VI following the same procedure as in the previous experiments. The model summary can be seen in Table 14. The only significant predictor is cue type (β = 2.45, se = 0.51, p $<$ 0.001).

We then compare the test data from Experiment VI to test data from Experiment II. Experiment II constitutes the best comparison since it also has no new conversation partners in test. The pattern type is the diminutive rather than the plural, but Experiment V provided little evidence that this would be a relevant dimension.

We merge the two test datasets to see whether test accuracy in Experiment VI (which has 18 training items) is better than in Experiment II (which has 6 training items), and whether this has any interaction with cue type (*gender* or *view*).

We stepwise fit a binomial mixed-effects regression model on the test data using response as an outcome variable and item presence in training, cue type, and participant age and gender as predictors, with a participant random intercept. The summary of the best model can be seen in Table 15[3].

We see that cue type has a strong positive effect on test accuracy. Training length matters. Participants who go through longer training are more accurate in the test phase. However, this effect is mostly carried by the *gender* cue: longer training is beneficial to those who are trained with the gender distinction. The effect plot can be seen in Figure 14.

If we tabulate good learners and poor learners across conditions and compare the results to Experiment II (which is similar in structure but has a shorter training phase), we find that the ratio of good learners increases with increased training (cf. Table 16 – there are more good learners in Exp VI than in Exp II.)

9.5 Discussion

Based on the results of experiments I-IV, we hypothesised that two types of information are available to participants in this experimental paradigm, the stem (the linguistic context) and the suffix (the non-linguistic context). If the non-linguistic context is salient, it is more readily available as a factor in variation. What remaines unclear is whether participants then decide to focus on the stem or the suffix (resulting in less or more success in the experiment) at random or based on specific learning strategies. Experiment VI shows that if we increase the training set, more participants are able to determine the relevant cue for the linguistic pattern (the non-linguistic context). This indicates that, if participants adopt a learning strategy (e.g. focussing on the prompt picture or on the stem), some of them are able to update it based on evidence that it does not yield good results. Increasing the amount of evidence available by lengthening the training phase enables more participants to modify their strategy and succeed at the task.

Age has no effect on test accuracy for participants with extended training length. This indicates that training can overcome the age effect. It is true, however, that the effect of age in our experiments is not robust with this sample size. This means that we have to be very cautious in interpreting the lack of an effect here. Ultimately, the relationship of training length and age could only be tested with a larger sample, which is outside the scope of this paper.

[3] The way the experimental platform assigns participants to conditions has a slight random element, and one of the conditions has considerably fewer participants in it. In order to make sure that our results are robust, we re-fit our model on a subset of the test data with 39 participants in each condition. The main effects did not change considerably.

10 SUMMARY

We have given a review of the literature to show that the non-linguistic context is extremely influential in learning linguistic constructions. Indeed, language use is shaped to a large degree by the social context. However, the link between contextual language learning and the observed structural complexity of social language use is far from completely understood.

We presented a series of artificial language learning experiments in which learning takes place in different contexts, which have different degrees of social-cognitive salience. The experiments were designed to investigate whether the relative social salience of contextual cues is relevant to learning a language pattern and whether this pattern is generalised to new lexical items and contexts. We hypothesised that participants would fare better at learning the link between the type of conversation partner and morphological pattern if the categorisation of conversation partners was socially salient. We also assumed that this salient link would be generalised to new items and new conversation partners.

We found that participants learned the association of two morphological patterns (a diminutive suffix and a plural suffix) with conversation partner identity or gender, much as they learned the linguistic pattern that we used as a baseline (a cooccurrence constraint between the stem and suffix vowels).

Successful learners of the contextual association generalised well to new items. However, learning contextual associations was overall rather difficult for the participants. There were substantial individual differences in learning. As in the survey of statistical learning in various domains by Siegelman and Frost (2015), we found that people vary in their individual ability to learn from training data – some people have high accuracy, and others perform worse.

The test data distributions generally showed two distinct modes, one for ‘good learners’ and one for ‘poor learners’. The adaptive tracking training enabled us to examine the differences between good learners and poor learners in more depth. Good learners finished the training phrase faster, suggesting that they identified and focused on the relevant cue better than poor learners did. However, even participants who were ‘good learners’ needed to make a number of mistakes to learn the pattern. The distributions of training trial counts for ‘good learners’ reveal that training took good learners longer than would be needed for a player who plays ideally. This means that each of them had to make at least a few mistakes before they learned the pattern. With the lengthened training phrase in Experiment VI, a greater number of opportunities to notice the relevant cue had the result that a greater number of participants responded to failure by readjusting their focus, ultimately patterning as good learners in the test phase.

An important result of these experiments is the relative success for different non-linguistic contextual dimensions. Social salience is very important. When the link between the conversation partner and the item appeared relatively accidental (side-facing, vs front-facing), the association was very difficult to learn. When the link was socially coherent and interpretable (conversation partner gender), the learning task was considerably easier (Experiments III, V, and VI). Participants learn relatively easily, for example, that a particular adult female calls a small *fen* a *fenwun*, whereas a different person – a male – calls it a *fentas*. Participants orient early to the contextual cue of gender, and easily generalise this both to new items and to other conversation partners (Experiment III). This is true even when little evidence of generality is actually given. That is, exposure to just one female partner saying *fenwun* (in two views) leads to the hypothesis that all females would prefer *fenwun* to *fentas*.

Another aspect of learning is the competition between the linguistic context and non-linguistic context. In Experiment I, where participants need to focus on the linguistic context (the prompt name), familiar items (ones seen in training) are chosen more accurately in test. In Experiments II and III, where participants need to focus on the non-linguistic context (along with the suffix), this effect is absent. This remains true even for the ‘poor learner’ participant group – those who did not seem to pin down the relevant contextual difference (conversation partner gender or spatial orientation).

In Experiment IV, we see that participants who focus on the suffix in test also finish training faster. This, in turn, supports the interpretation that the two types of information (stem and suffix) are competing with each other. Concentrating on the suffix is the key to success. Experiment V shows that this learning strategy is robust (applies for learning both the plural and the diminutive) while Experiment VI shows that the choice of stem or suffix as the main locus of attention is not fixed. With increased training, more participants figure out the relevant dimension and respond like ‘good learners’ in the test phase.

These results can be compared with the results of Lleras and Von Mühlenen (2004). They find that, in a learning experiment, where contextual cues correlate with tasks, participants that employ an ‘active’ searching strategy, and focus on the task itself, do not rely on contextual cues. The participants in our experiments follow an analogous pattern. Those who focus on the stem ignore social contextual cues. For the baseline Experiment I, the social contextual information is irrelevant and focusing on the stem would lead to success; but in the other experiments, focusing on the stem would cause the participant to overlook the information that is actually relevant to the task. The interesting point is that *whether* they focus on the stem or the suffix depends on the kind of social contextual cue present. They are more likely to rely on the social contextual cue if it is salient.

Taken together, the results of these experiments provide solid evidence that adults are able to learn contextual meaning, and that they orient more toward contextual information that is socially salient and relevant than to contextual variation that appears accidental.

11 GENERAL DISCUSSION

The focus of this article is learning associations between a morphological pattern and a non-linguistic context. The main question is how the social salience of the context influences success in learning.

We contrasted the learning of two cues, one of which is socially salient (*gender*) and one of which is not (*view*), showing that the former is learned more easily than the latter. As we note in Section 3, the perceptual differences between the images are unlikely to affect their categorisation.

Of course, the forced-choice paradigm has its limitations. When we say that participants were better at learning the *gender* cue, this needs to be interpreted within the context of the task. We do not know whether they learned it well enough to produce it unprompted, for example, nor do we know whether, outside of a forced-choice task, they would have preferred some unknown other response. The positive side of a forced-choice paradigm is that our results are easy to interpret statistically. But the results do open up further questions about how the results would pattern if a free-response paradigm was used. As we only contrast gender and view, a further question is whether these two conditions may vary on unknown dimensions other than salience. Further research using other images and other contrasts is therefore still required.

Social salience has a top-down effect. Prior experience teaches us that some differences are more important than others, and we pay more attention to these in linguistic categorisation. The way we see the world, then, has a strong influence on our language use, resulting in the complex structures of indexicality discussed by sociolinguists on the population level. This article provides evidence that this influence is present on an individual level. The social salience of the images is likely to rely on more than ”gender”, but it remains the core manipulation that participants react to. The manipulation appears to provide a very reliable effect, despite the simplicity of the experimental paradigm.

Our results indicate that participants give more accurate answers when they recognise the relevant distinction (e.g. female/male in the *gender* condition). This is broadly analogous to explicit social stereotypes in terms of recognising both the pattern and the context, as well as the connection between the two. At the same time, it can also be extended to cases in which either the context or the pattern is recognised and negotiated explicitly. The former is typical of most cases of social-indexical variation, in which we know our conversation partner's principal characteristics. The latter is typical of word patterns in particular, such as the choice between the formal and the informal second person conjugation in French and German, and the dialect-specific vocabularies of English, German, French, and many other national languages.

Experiment I, as well as the combined analysis of Experiments II, III, and V, showed that older participants were more successful at both morpho-phonological and socio-indexical learning.

The age effect may arise because prior experience, increasing throughout the lifespan, has a beneficial effect on learning tasks, as proposed by Ramscar et al. (2014). As older participants were better at learning all associations presented (including the un-natural *view* association), this cannot be viewed as an effect of increased experience with socially relevant distinctions. Rather, it would have to be interpreted as an effect of increased general experience with learning socio-indexical and linguistic associations. The age effect may also arise in some way from the specific nature of our task. For example, if participants select the wrong answer, they get feedback. This feedback could provide them with information that are orienting to the wrong cue. There is some evidence that older participants make better use of feedback, especially in situations in which they are initially uncertain (Metcalfe et al., 2016).

It is important to note that the age range of our participants is restricted, when considered in the context of the literature on ageing. Only one of our 498 participants (with age data available) is over 70. The considerable literature on cognitive decline in ageing across a range of psychological tasks compares younger and older adults (usually over 70), with an assumption that speakers in the middle (i.e. 40-60) fall somewhere in between (Lachman, 2004). While the middle-aged group tends to be less studied, there are at least some studies which show improvement from younger adults to middle age, before declining again in older adults. Tasks where such an effect has been reported include everyday problem-solving (Thornton et al., 2013) and social problem-solving (D'Zurilla et al., 1998).

It is interesting to note that we see no age effect for the task with extended training (Experiment VI). This may suggest that, whatever the root of the age effect, additional practice can neutralise the benefits of increased age in this task.

Our results indicate a major role for social salience in the acquisition of contextual meaning in morphology. In our task, the contextual information is associated with the morphological pattern in the cognitive representation, and influences recall and generalization. Whether or not participants are overtly aware of the association, it is sufficient to nudge them in the correct direction in a two-way forced choice test. The fact that the gender-dependent association is learned better than an accidental association, and that performance slightly improves with age, reveals the role of prior knowledge and expectation about what aspects of the context may potentially be relevant. Many aspects of language vary according to the gender of the conversation partner, and in the participants' prior sociolinguistic learning, the gender of the conversation partner will have been relevant many times. Foulkes (2010) hypothesises that some types of indexical properties should be more readily transmitted than others, based on the frequency with which they have been relevant in individuals' past experience. He identifies gender as one of the earliest learned socio-indexical associations. Children as young as 6 months, for example, preferentially match sex-cued voices and faces (Walker-Andrews et al., 1991). It is likely this considerable prior experience that facilitates a ready generalisation across conversation partners.

12 CONCLUSION

Our paradigm demonstrates differences in adult learning of socially salient versus accidental non-linguistic contextual cues. It also reveals a number of questions about the way we learn contextual associations of higher level linguistic structures. Does a varying non-linguistic context aid the learning of a linguistic pattern? Do we learn the diminutive more easily, for example, if we are exposed to more types of conversation partners who use it? What is the effect of attention to particular linguistic patterns and non-linguistic contexts? Does variance in a non-linguistic difference that we explicitly attend to aid language learning? Finally, amongst socially salient non-linguistic cues, are some easier to learn than others? Is it easier to learn the association of a linguistic pattern with gender, for example, than with age? These questions remain to be answered by follow-up research.

Our controlled laboratory experiments are, of course, still many worlds apart from the type of complex socio-contextual learning and generalisation that occurs in every day interaction and language acquisition. However they do provide some first steps towards shedding some light on the complex cognitive mechanisms that must be at play in such learning. Whether an associative pattern is attended to, learned and recreated by a speaker will be affected by a range of factors - including who that individual is, how socially salient the relevant context is, and how much the learner is exposed to that association. Our experiments have shown that individual variability in individual listeners, the salience of socio-contextual associations, and differing patterns of exposure, likely all play some role in affecting socio-contextual learning in morphology.

DISCLOSURE/CONFLICT-OF-INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

P. R. designed and ran the experiments, performed statistical analysis, and wrote up the results. J. H. and J. B. P. contributed to the design, provided feedback on the experiments and analysis and contributed to writing up the manuscript.

ACKNOWLEDGMENTS

The authors would like to thank Chun Liang Chan, Viktória Papp, Alex Schumacher, Márton Sóskuthy, Clay Beckner, Patrick LaShell and Steven Franconeri.

Funding: This project was made possible through the support of a grant to Northwestern University from the John Templeton Foundation (Award ID 36617). The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of the John Templeton Foundation. J. H. was also supported by a Rutherford Foundation Grant (Grant Number E5909).

REFERENCES

Altmann, E. G., Pierrehumbert, J. B., and Motter, A. E. (2011). Niche as a determinant of word fate in online groups. *PloS one* 6, e19009

Baayen, R. H., Hendrix, P., and Ramscar, M. (2011). Sidestepping the combinatorial explosion: Towards a processing model based on discriminative learning. In *Empirically examining parsimony and redundancy in usage-based models, LSA workshop*

Bates, D., Maechler, M., and Bolker, B. (2012). lme4: Linear mixed-effects models using s4 classes. R package

Bauer, L. (1997). Evaluative morphology: in search of universals. *Studies in Language* 21, 533–575

Baxter, G. J., Blythe, R. A., Croft, W., and McKane, A. J. (2009). Modeling language change: An evaluation of Trudgill's theory of the emergence of New Zealand English. *Language Variation and Change* 21, 257–296

Becker, M., Ketrez, N., and Nevins, A. (2011). The surfeit of the stimulus: analytic biases filter lexical statistics in Turkish laryngeal alternations. *Language* 87, 84–125

Beckner, C., Rácz, P., Brandstetter, J., Hay, J., and Bartneck, C. (2016). Participants conform to humans but not to humanoid robots in an English past tense formation task. *Journal of Language and Social Psychology* 35, 158–79. doi:10.1177/0261927X15584682

Brooks, P. J., Kwoka, N., and Kempe, V. (2016). Distributional effects and individual differences in l2 morphology learning. *Language Learning*

Campbell-Kibler, K. (2011). Intersecting variables and perceived sexual orientation in men. *American Speech* 86, 52–68

Cheshire, J. (2002). Sex and gender in variationist research. *Handbook of Language Variation and Change* , 423–443

Chun, M. M. and Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive psychology* 36, 28–71

Coupland, J., Coupland, N., and Giles, H. (1991). Accommodation theory. communication, context and consequences. *Contexts of Accommodation. Cambridge & Paris: Cambridge University Press & Éditions de la maison des sciences de lhomme* , 1–68

Crump, M. J., McDonnell, J. V., and Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS one* 8, e57410

Denton, S. E., Kruschke, J. K., and Erickson, M. A. (2008). Rule-based extrapolation: A continuing challenge for exemplar models. *Psychonomic bulletin & review* 15, 780–786

Dixon, R. M. W. (1980). *The languages of Australia* (Cambridge University Press)

Docherty, G. J., Langstrof, C., and Foulkes, P. (2013). Listener evaluation of sociophonetic variability: Probing constraints and capabilities. *Linguistics* 51, 355–380

Drager, K. (2010). Sociophonetic variation in speech perception. *Language and Linguistics Compass* 4, 473–480

D'Zurilla, T. J., Maydeu-Olivares, A., and Kant, G. L. (1998). Age and gender differences in social problem-solving ability. *Personality and individual differences* 25, 241–252

Eckert, P. (2000). *Linguistic variation as social practice: The linguistic construction of identity in Belten High* (Wiley-Blackwell)

Eckert, P. (2008). Variation and the indexical field1. *Journal of sociolinguistics* 12, 453–476

Eckert, P. (2012). Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual review of Anthropology* 41, 87–100

Fagyal, Z., Swarup, S., Escobar, A. M., Gasser, L., and Lakkaraju, K. (2010). Centers and peripheries: Network roles in language change. *Lingua* 120, 2061–2079

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, 17897–17902

Foulkes, P. (2010). Exploring social-indexical knowledge: A long past but a short history. *Laboratory Phonology* 1, 5–39

Foulkes, P. and Docherty, G. (2006). The social life of phonetics and phonology. *Journal of Phonetics* 34, 409–438

Foulkes, P. and Hay, J. B. (2015). 13 the emergence of sociophonetic structure. *The handbook of language emergence* 87, 292

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models* (Cambridge University Press)

German, J. S., Carlson, K., and Pierrehumbert, J. B. (2013). Reassignment of consonant allophones in rapid dialect acquisition. *Journal of Phonetics* 41, 228 – 248

Gluck, M. A. and Myers, C. E. (2001). *Gateway to memory: An introduction to neural network modeling of the hippocampus and learning* (MIT Press)

Godden, D. R. and Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of psychology* 66, 325–331

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science* 13, 431–436

Goujon, A., Didierjean, A., and Thorpe, S. (2015). Investigating implicit statistical learning mechanisms through contextual cueing. *Trends in cognitive sciences* 19, 524–533

Grossberg, S. (1987). Competitive learning: From interactive activation to adaptive resonance. *Cognitive science* 11, 23–63

Hay, J. and Walker, A. (2013). Skewed experience with words affects lexical access patterns. In *Variation and Language Processing (VALP) 2, Christchurch, January*

Hay, J. B. and Drager, K. (2007). Sociophonetics. *Annual Review of Anthropology* 36, 89–103

Hay, J. B. and Drager, K. (2010). Stuffed toys and speech perception. *Linguistics* 48, 865–892

Henderson, L., Devine, K. W., and Gaskell, M. (2015). When the daffodat flew to the intergalactic zoo: Off-line consolidation is critical for word learning from stories. *Developmental Psychology* 51, 406–417

Horton, W. S. and Gerrig, R. J. (2002). Speakers experiences and audience design: Knowing when and knowing how to adjust utterances to addressees. *Journal of Memory and Language* 47, 589–606

Horton, W. S. and Gerrig, R. J. (2005). The impact of memory demands on audience design during language production. *Cognition* 96, 127–142

Itti, L., Koch, C., Niebur, E., et al. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence* 20, 1254–1259

Johnson, K., Strand, E. A., and D'Imperio, M. (1999). Auditory–visual integration of talker gender in vowel perception. *Journal of Phonetics* 27, 359–384

Jurafsky, D. (1993/2012). Universals in the semantics of the diminutive. In *Annual Meeting of the Berkeley Linguistics Society*. vol. 19

Kraljic, T., Brennan, S. E., and Samuel, A. G. (2008a). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition* 107, 54–81

Kraljic, T. and Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic bulletin & review* 13, 262–268

Kraljic, T., Samuel, A. G., and Brennan, S. E. (2008b). First impressions and last resorts. How listeners adjust to speaker variability. *Psychological science* 19, 332–338

Kruisinga, E. (1942). *Diminutieve en affektieve suffixen in de Germaanse talen* (Noord-hollandsche uitgevers maatschappij)

Labov, W. (2001). *Principles of linguistic change: Social factors* (Wiley-Blackwell)

Lachman, M. E. (2004). Development in midlife. *Annu. Rev. Psychol.* 55, 305–331

Langstrof, C. (2014). *Sociophonetic learning in L1 and L2* (Habilitation, University of Freiburg)

Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Perception & psychophysics* 63, 1279–1292

Leung, J. H. and Williams, J. N. (2012). Constraints on implicit learning of grammatical form-meaning connections. *Language Learning* 62, 634–662

Leys, C., Ley, C., Klein, O., Bernard, P., and Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology* 49, 764–766

Lleras, A. and Von Mühlenen, A. (2004). Spatial context and top-down strategies in visual search. *Spatial Vision* 17, 465–482

Mathworks (2016). *MATLAB version 9.1.0.441655 (R2016b).* The Mathworks, Inc., Natick, Massachusetts

Maye, J., Aslin, R. N., and Tanenhaus, M. K. (2008). The weckud wetch of the wast: Lexical adaptation to a novel accent. *Cognitive Science* 32, 543–562

Metcalfe, J., Casal-Roscum, L., Radin, A., and Friedman, D. (2016). On teaching old dogs new tricks. *Psychological Science* , doi:10.1177/0956797615597912

Milroy, J. and Milroy, L. (1993). Mechanisms of change in urban dialects: The role of class, social network and gender. *International Journal of Applied Linguistics* 3, 57–77

Milroy, L. (1980). *Language and social networks* (Oxford University Press)

Munro, R., Bethard, S., Kuperman, V., Lai, V. T., Melnick, R., Potts, C., et al. (2010). Crowdsourcing and language studies: The new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (Association for Computational Linguistics), 122–130

Needle, J. N., Pierrehumbert, J. B., and Hay, J. B. (2015). Effects of pseudomorphology on the wordlikeness of pseudowords. In *Architectures and Mechanisms for Language Processing'. 3 - 5 September. University of Malta Valletta*

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14, 700

Pele, O. and Werman, M. (2010). The quadratic-chi histogram distance family. In *European conference on computer vision* (Springer), 749–762

Pierrehumbert, J. B. (2006). The next toolkit. *Journal of Phonetics* 34, 516 – 530

Pierrehumbert, J. B., Bent, T., Munson, B., Bradlow, A. R., and Bailey, J. M. (2004). The influence of sexual orientation on vowel production. *J. Acoust. Soc. Am.* 116, 1905–1908

Pierrehumbert, J. B., Stonedahl, F., and Daland, R. (2014). A model of grassroots changes in linguistic systems. *arXiv preprint arXiv:1408.1985*

Preston, D. R. (1996). Whaddayaknow?: The modes of folk linguistic awareness. *Language awareness* 5, 40–74

Qian, T., Jaeger, T., and Aslin, R. N. (2014). Implicit learning in a non-stationary environment: Knowing when it's Groundhog Day. Ms, University of Rochester, Retrieved: 2014-03-17 19:54:18

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria

Rácz, P. (2013). *Salience in Sociolinguistics: A Quantitative Approach* (De Gruyter)

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., and Baayen, H. (2014). The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in cognitive science* 6, 5–42

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological review* 85, 59

Roberts, G. (2008). Language and the free-rider problem: An experimental paradigm. *Biological Theory 3* 2, 174–183

Rost, G. C. and McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental science* 12, 339–349

Säily, T. and Suomela, J. (2009). Comparing type counts: the case of women, men and *-ity* in early English letters. In *Corpus linguistics: Refinements and reassessments*, eds. A. Renouf and A. Kehoe (Amsterdam: Rodopi). 87–109

Schumacher, R. A., Pierrehumbert, J. B., and LaShell, P. (2014). Reconciling inconsistency in encoded morphological distinctions in an artificial language. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society. Cognitive Science Society*

Siegelman, N. and Frost, R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of memory and language* 81, 105–120

Silverstein, M. (2009). The limits of awareness. In *Linguistic Anthropology: A reader*, ed. A. Duranti (Oxford: Blackwell), vol. 1. 382–401

Smith, S. M. and Vela, E. (2001). Environmental context-dependent memory: A review and meta-analysis. *Psychonomic bulletin & review* 8, 203–220

Soliz, J. and Giles, H. (2014). Relational and identity processes in communication: A contextual and meta-analytical review of communication accommodation theory. *Communication yearbook* 38, 106–143

Tagliamonte, S. A. and Roeder, R. V. (2009). Variation in the English definite article: Socio-historical linguistics in t'speech community. *Journal of Sociolinguistics* 13, 435–471

Thornton, W. L., Paterson, T. S., and Yeung, S. E. (2013). Age differences in everyday problem solving: The role of problem context. *International Journal of Behavioral Development* 37, 13–20

Van der Zande, P., Jesse, A., and Cutler, A. (2014). Cross-speaker generalisation in two phoneme-level perceptual adaptation processes. *Journal of Phonetics* 43, 38–46

Von Ahn, L. (2006). Games with a purpose. *Computer* 39, 92–94

Walker, A. and Hay, J. B. (2011). Congruence between 'word age' and 'voice age' facilitates lexical access. *Laboratory Phonology* 2, 219–237

Walker-Andrews, A. S., Bahrick, L. E., Raglioni, S. S., and Diaz, I. (1991). Infants' bimodal perception of gender. *Ecological Psychology* 3, 55–75

Wells, J. C. (1982). *Accents of English*, vol. 1 (Cambridge University Press)

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York)

Yu, C. and Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science* 18, 414–420

TABLES

fek	ran	wek	fal
pel	ral	tek	rak
tas	fan	wen	fes

Table 1 Stimuli set, Experiment I

Formula: correct $\sim$ item in training + age + (1 + item in training \| participant)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.17	1.17	-1.85	0.06
item in training = TRUE	1.14	0.33	3.49	0.00
age	0.11	0.04	3.05	0.00

Table 2 Best model summary for Experiment I

fek	rik	wuk	fal
pel	ril	tol	rul
wan	fen	wun	tas
fis	tos		

Table 3 Stimuli set, Experiment II

Formula: correct $\sim$ age + cue type + (1 \| participant)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.51	0.63	-2.40	0.02
age	0.06	0.02	3.21	0.00
cue type = gender	1.09	0.34	3.20	0.00

Table 4 Best model summary for Experiment II

	view cue	gender cue
good learner	5	25
poor learner	40	27

Table 5 Good learners and poor learners across cue type, Experiment II

Formula: correct ∼ cue type + (1 | participant)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.28	0.31	0.90	0.37
cue type = gender	1.68	0.44	3.81	0.00

Table 6 Best model summary, Experiment III

	view cue	gender cue
good learners	8	25
poor learners	41	23

Table 7 ‘Good’ learners across cue type, Experiment III

	Training		Test	
Exp II	base base+correct suffix base+incorrect suffix	tas tasrul taspel	base base+correct suffix base+incorrect suffix	tas tasrul taspel
Exp IV	base base+correct suffix base+incorrect suffix	tas tasrul taspel	(base not visible) incorrect base + correct suffix correct base + incorrect suffix	 fenrul taspel

Table 8 Example stimuli, Experiment II vs Experiment IV

correct base ∼ cue type + (1 | participant)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.86	0.44	4.27	0.00
cue type = gender	-2.81	0.62	-4.54	0.00

Table 9 Best model summary, Experiment IV

[ht]

correct ∼ cue type + (1 | participant)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.77	0.33	2.33	0.02
cue type = gender	0.84	0.46	1.83	0.07

Table 10 Best model summary, Experiment V

[ht]

correct ∼ cue type + (1 | participant)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.50	0.23	2.22	0.03
cue type = gender	1.29	0.32	4.03	0.00

Table 11 Best model summary, Experiments III and V

correct ∼ age + cue type + (1 | participant)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.51	0.43	-1.20	0.23
cue type = gender	1.23	0.25	5.01	0.00
age	0.03	0.01	2.45	0.01

Table 12 Best model summary, Experiments II, III, and V

fos	ruk	wik	ril
fol	pil	fel	tos
fon	tang	rong	fok
tel	fek	rel	tas
feng	fong	ros	wis
wal	tal	rek	pek
pung	fus	tol	rik
wun	rak	ren	ral
tus	wus	rok	tok

Table 13 Stimuli set, Experiment VI

correct ∼ cue type + (1 | participant)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.18	0.38	3.10	0.00
cue type = gender	2.45	0.51	4.77	0.00

Table 14 Best model summary, Experiment VI

Formula: correct ∼ training length * cue type + (1 | participant)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.64	0.34	4.82	0.00
training length = 18 items	1.94	0.46	4.24	0.00
cue type = gender	-0.86	0.48	-1.78	0.08
training length = 18 items:cue type = gender	-1.56	0.67	-2.34	0.02

Table 15 Best model summary, Experiments II and VI

	Experiment II	(6 training items)	Experiment VI	(18 training items)
	view cue	gender cue	view cue	gender cue
good learners	5	**25**	12	**41**
poor learners	40	**27**	28	**15**

Table 16 ‘Good’ learners across cue type, Experiments II and VI

FIGURES

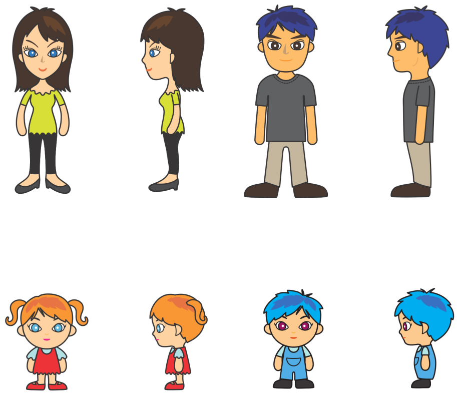

Figure 1. The eight conversation partner images used in the experiments

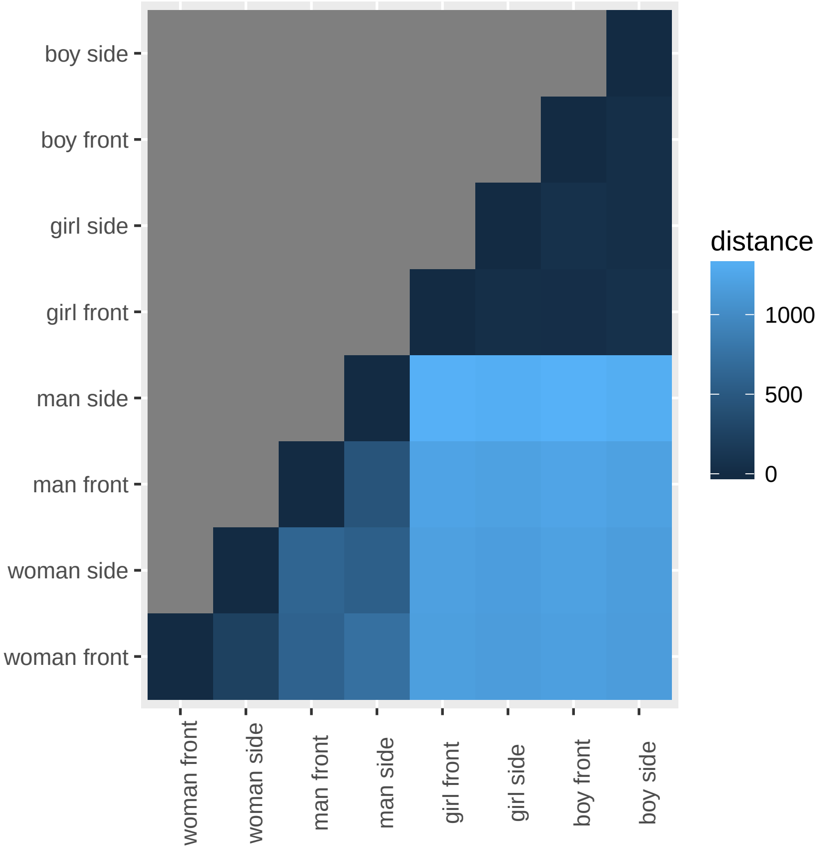

Figure 2. Distances between the eight conversation partner images

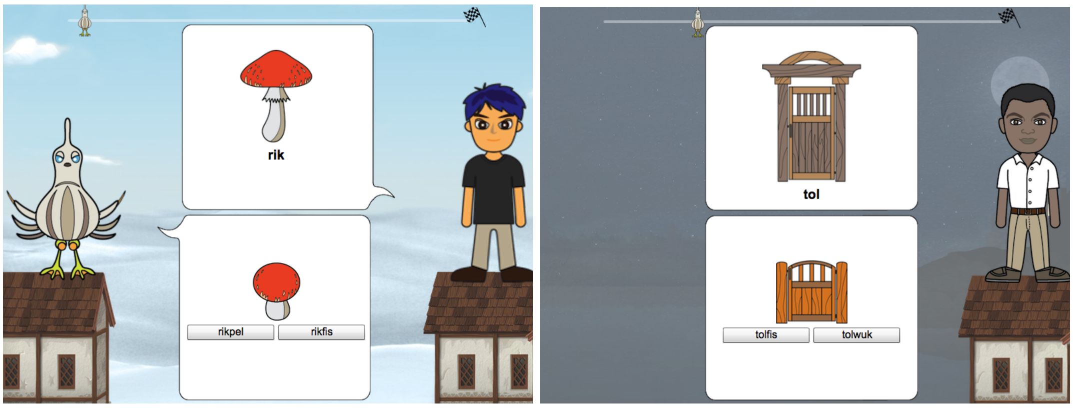

Figure 3. Left: The in-game set-up of the training phase in all our experiments. The player is on the left, the conversation partner is on the right. The query is in the speech bubble that belongs to the conversation partner. The response choice buttons are in the speech bubble that belongs to the player. One of the answers is correct, the other one is wrong. In Experiment I, the correct answer depends on the stem vowel of the prompt. In the rest of the experiments, it depends on the conversation partner (as in this example). Right: The test phase. The visuals separate it from training.

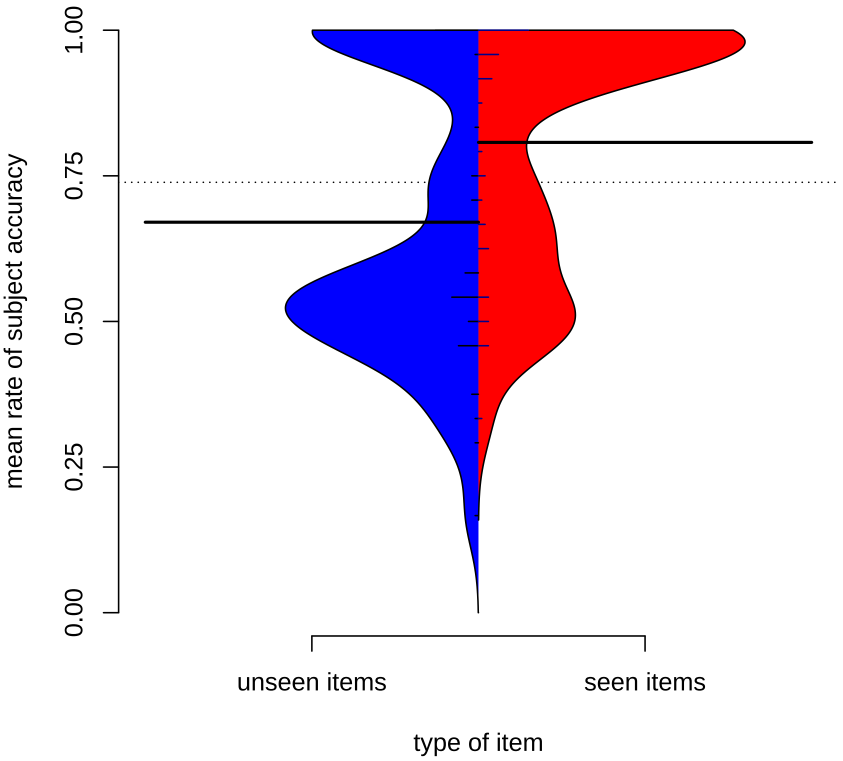

Figure 4. Distributions of participant responses on previously seen and unseen test items, Experiment I. Black horizontal bars show the mean accuracy for each set of items. The dotted line shows the overall mean. Small horizontal lines show individual values; longer if multiple individuals have the same average.

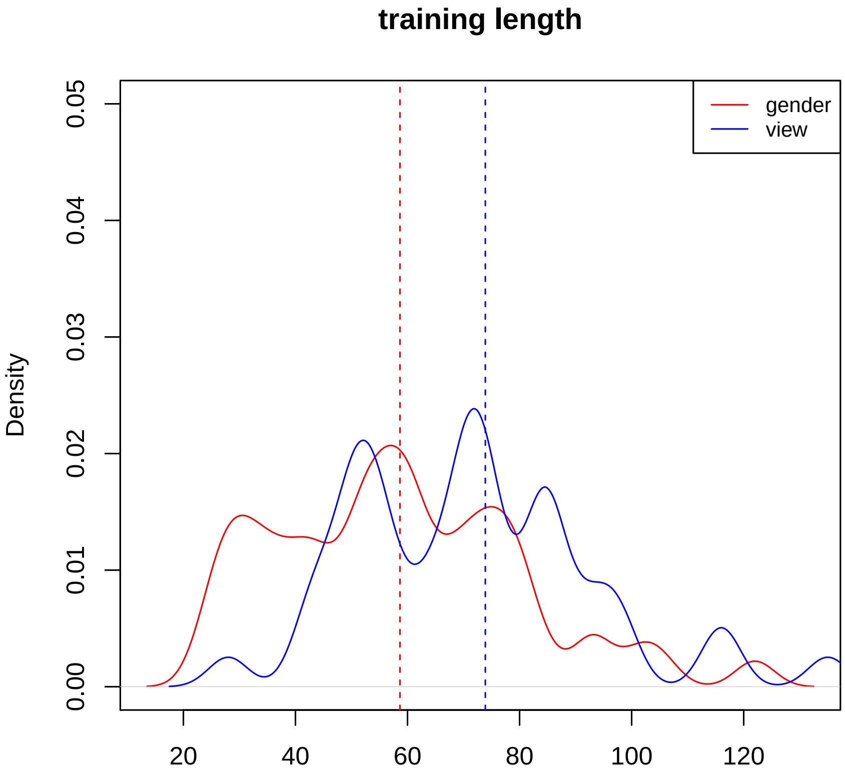

Figure 5. Distributions of training trial counts for the two conditions for all participants, Experiment II.

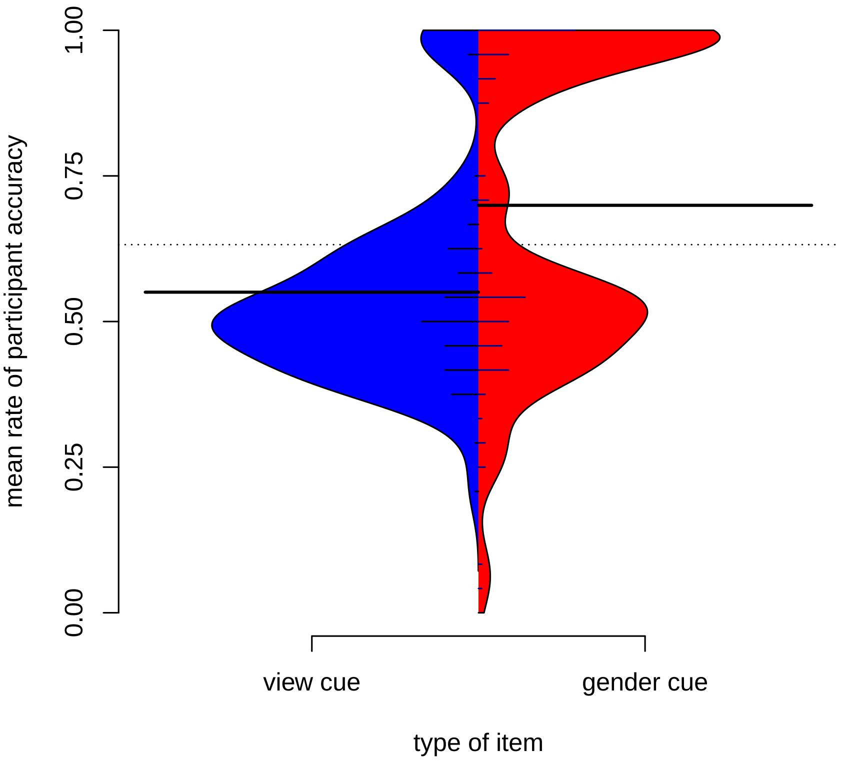

Figure 6. Distributions of participant responses with different contextual cues, Experiment II. Black horizontal bars show the mean accuracy for each set of items. The dotted line shows the overall mean. Small horizontal lines show individual values; longer if multiple individuals have the same average.

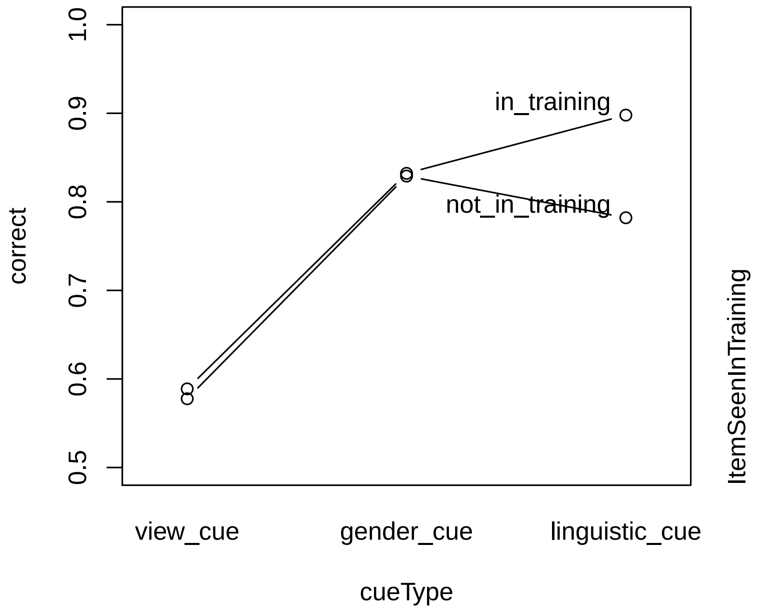

Figure 7. Effect of cue type (view cue, gender cue, or linguistic cue) and item presence in training in the test data for Experiments I and II. Results of mixed effects model on combined data set.

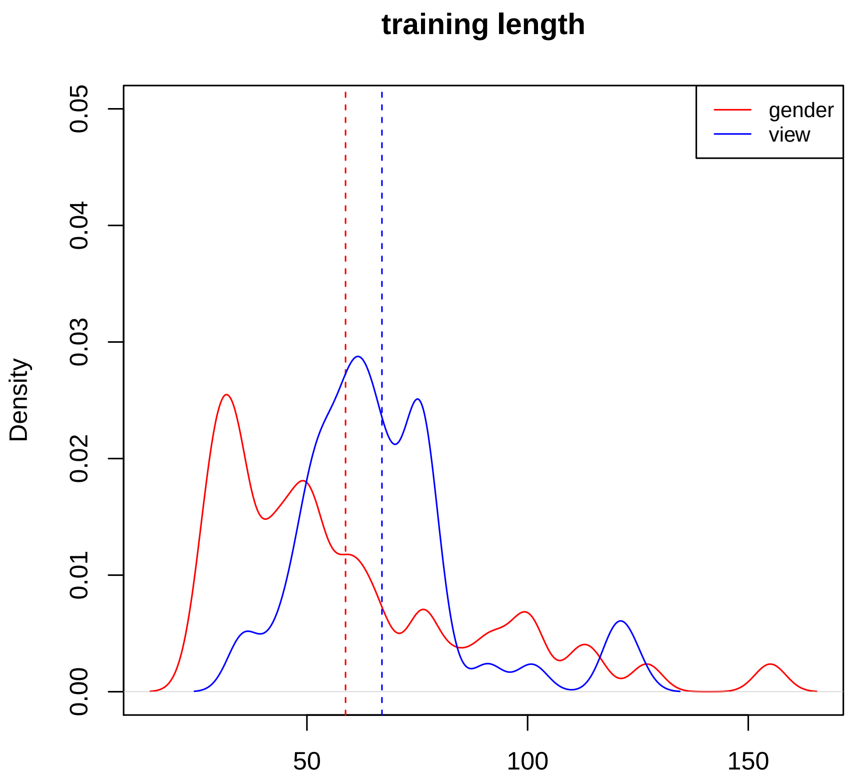

Figure 8. Distributions of training trial counts for the two conditions for all participants, Experiment III.

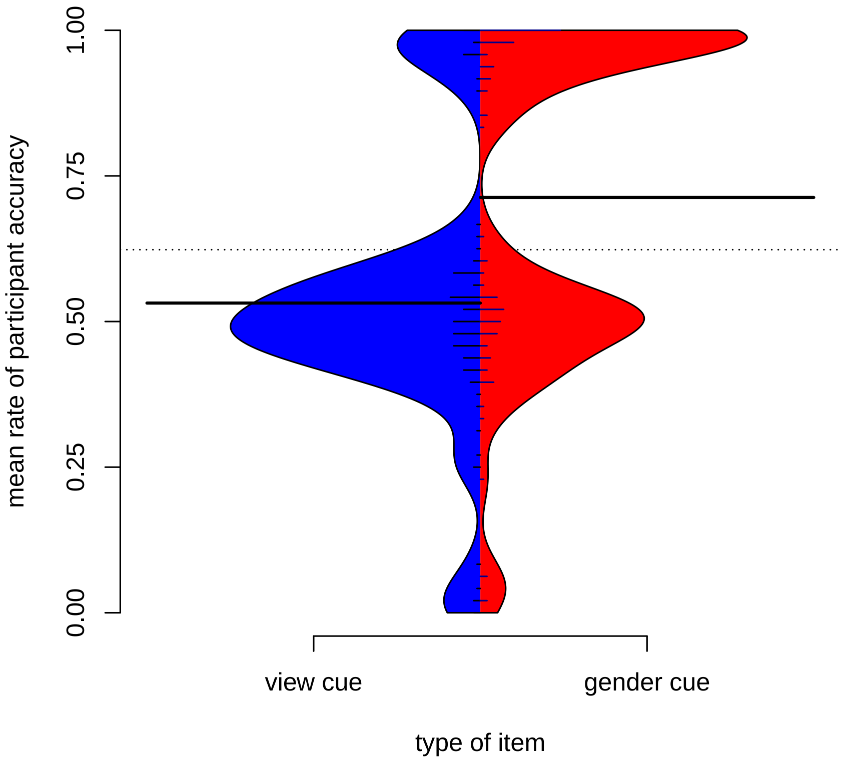

Figure 9. Distributions of participant responses on test items in the *view* and *gender* conditions, Experiment III. Black horizontal bars show the mean accuracy for each condition. The dotted line shows the overall mean. Small horizontal lines show individual values; longer if multiple individuals have the same average.

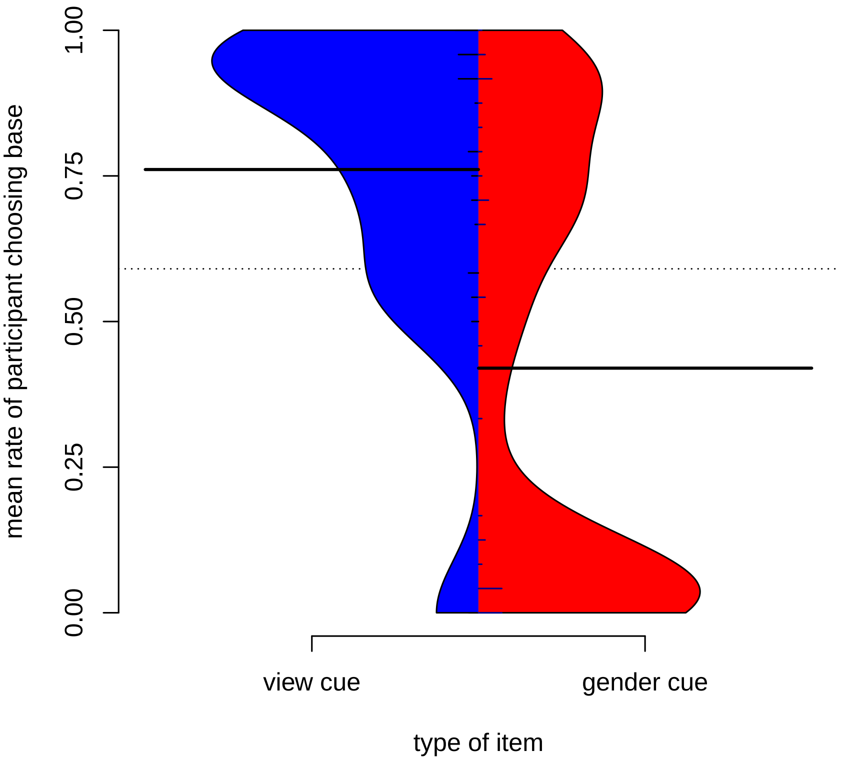

Figure 10. Distributions of participants picking right base plus wrong suffix (1) versus wrong base plus right suffix (0) in the two conditions, Experiment IV. Black horizontal bars show the mean rate of 'base' preference for each condition. The dotted line shows the overall mean. Small horizontal lines show individual values; longer if multiple individuals have the same average.

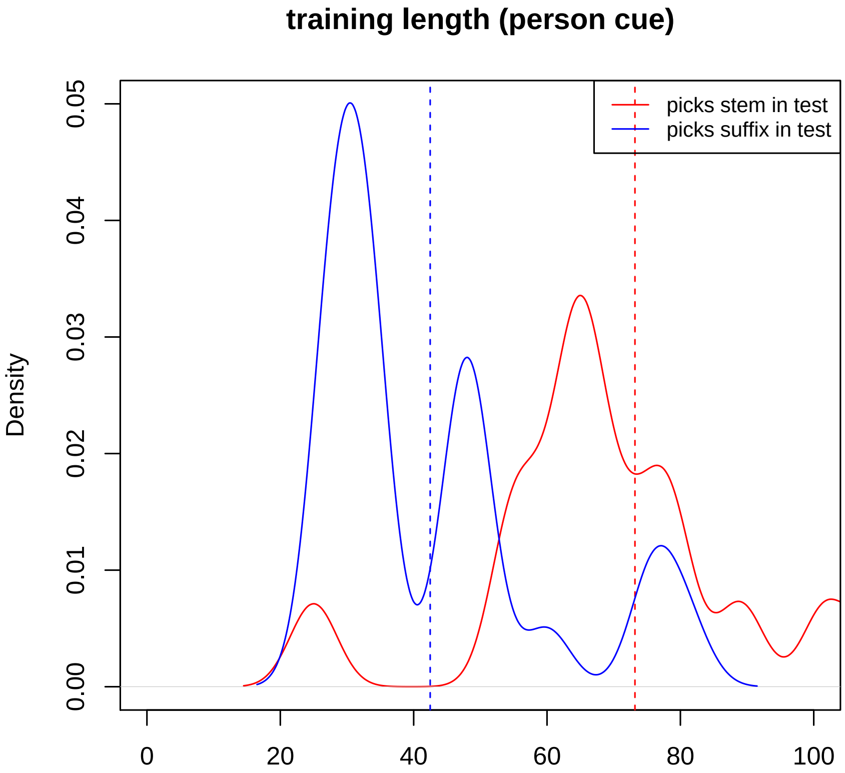

Figure 11. Distributions of training trial counts for ‘object’ and ‘people’ participants, Experiment IV.

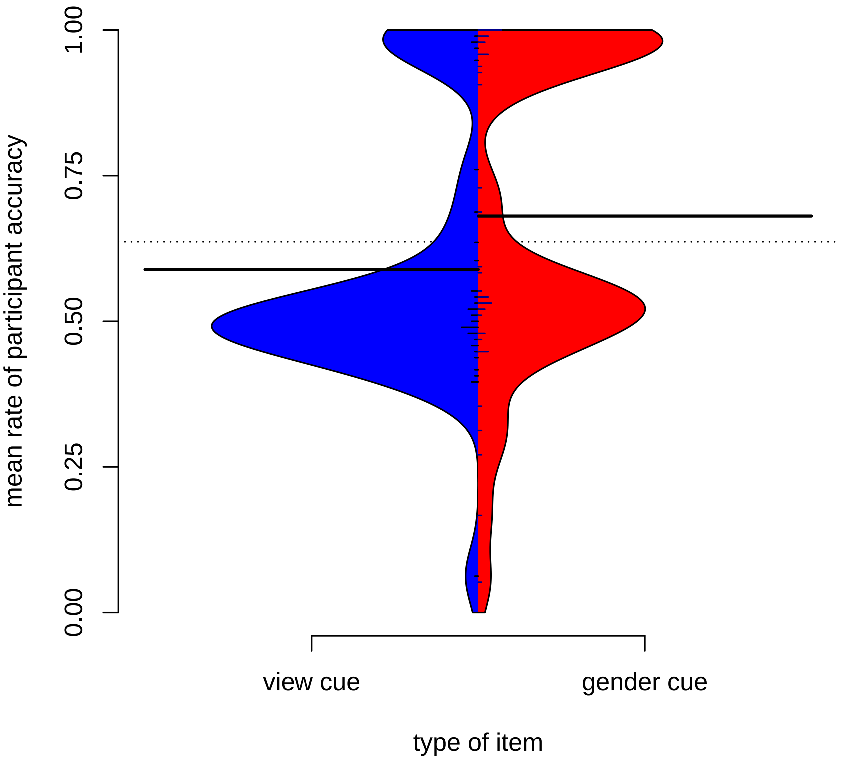

Figure 12. Distributions of participant responses on test items in the *view* and *gender* conditions, Experiment V. Black horizontal bars show the mean accuracy for each condition. The dotted line shows the overall mean. Small horizontal lines show individual values; longer if multiple individuals have the same average.

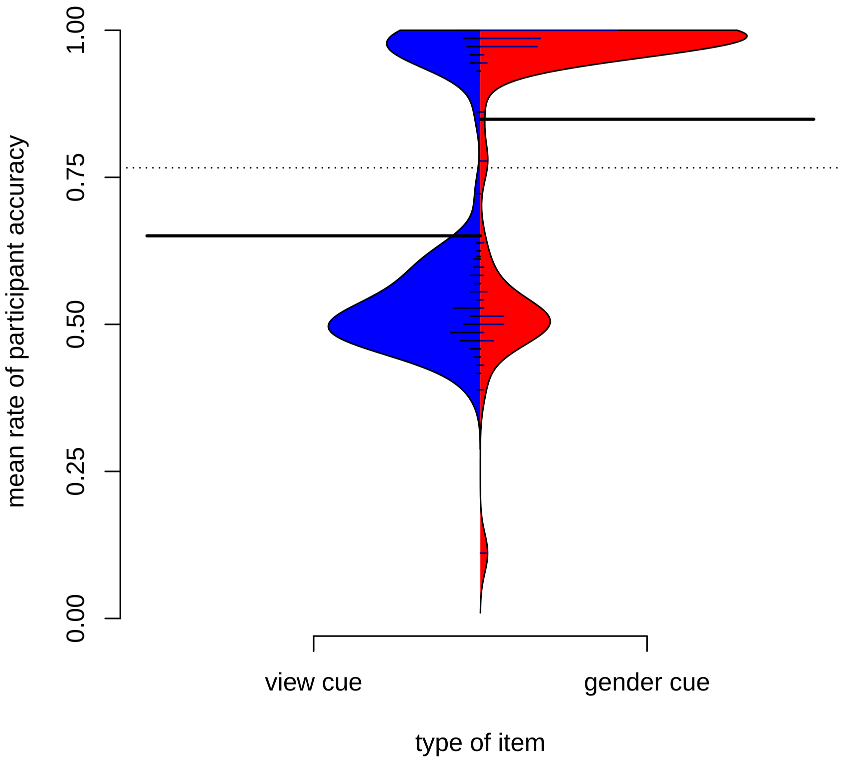

Figure 13. Distributions of participant responses on test items in the *view* and *gender* conditions, Experiment VI. Black horizontal bars show the mean accuracy for each condition. The dotted line shows the overall mean. Small horizontal lines show individual values; longer if multiple individuals have the same average.

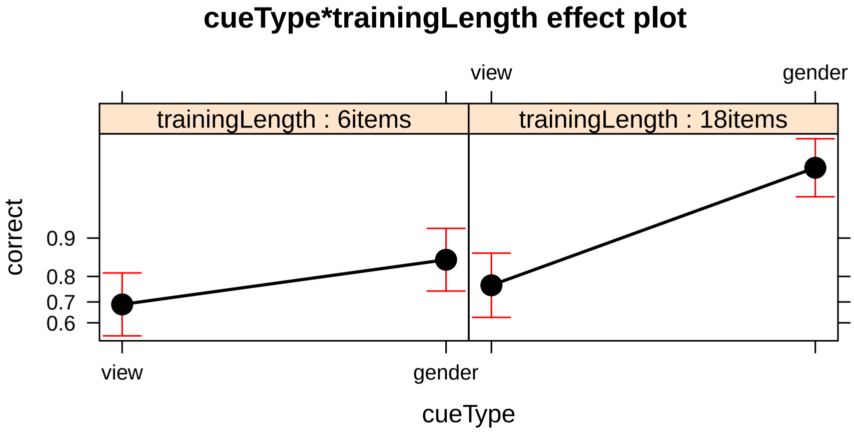

Figure 14. Interaction term from the regression model: Participant trained with the gender cue benefit from longer training in Experiment VI.